East of Walden Two

A Defence of Libertarian Freedom

Peter Rauschenberger

Contents

Many thanks to

Howard Robinson, my supervisor, for his generosity, kindness and rigour throughout the years of writing my PhD thesis, which is a shorter version of the present work, for which I am very grateful, Nenad Mišćević, my former supervisor, who was a great source of inspiration and encouragement to take philosophy seriously, Simon Saunders for supervising my progress during my year in Oxford in 2003-4, Martha Klein for the class on freedom and blameworthiness in 2004, and for commentary on an early version of chapter 8, Harvey Brown for the class on special relativity in 2004, and for an important discussion on determinism, Gábor Takács, for explanations and discussions on the consistent histories approach to quantum mechanics and the EPR-Bell experiment, András Szigeti, Andrea Csillag and Linda Lázár for intellectual companionship and for reading various drafts and giving many valuable comments.

1 Introduction: The Problem of Alternatives and Control

Alternatives and control

What I take to be our natural, commonsense conception of freedom is the conjunction of the following two ideas.

1) There are moments in our lives when more than one future is compatible with everything that has taken place that far. This includes all physical facts and events, and, if there are non-physical facts and events, them as well, inside and outside us. So this includes, among other things, a full description of our state at the given moment, physical and mental, and a full description of our history, physically and mentally speaking. Now, with this totality of past and present facts and events being as they are, it is objectively possible for the history of the universe, with our lives in it, to continue more than one way.

2) We control which of the possible continuations will actually take place.

This is the conjunction of two kinds of liberty, really, freedom in the negative sense, i.e., from determination by what has already been laid down, and freedom in the positive sense, i.e., controlling what is to come.

But this conjunction is problematic. On the face of it, the two conjuncts do not seem to sit well with each other.

It is natural to assume that we control which of the possible future courses of events is to occur by acting one way rather than the other. It is equally natural to assume that controlling how we act is achieved by the determination of our action by an intention or will to act that way, which is a state of mind.

But then the first of the two conjuncts, idea 1, is not true *of our action*, because it is determined by a prior state of mind. By the time our action occurs, there are no alternatives, since the will is already set. If there were alternatives even though the will is set, it is hard to see how the will would control the action.

I take it to be uncontroversial that freedom does not require alternatives relative to what is willed if the will is already set. But then it is natural to demand that ideas 1 and 2 apply to the event of our will's being set.

At this level the apparent conflict between control and alternatives is not so easily escaped. It is possible that our control over our will consists in its being determined by (some part of) what we are and how we are at the time. But then we have no alternatives. It is also possible that what we are and how we are underdetermines our will, so there are alternatives. But then it is hard to see how we manage control. *Prima facie* absolutism about negative liberty seems to rob us of all means of exercising positive freedom.

The libertarian conception of control and freedom

Perhaps it is only hard and not impossible to see how we manage control if the will is underdetermined. I will call the idea that control of the will can be achieved without the causal determinatedness of the will by facts or events that previously obtained the libertarian conception of control.

Following the received terminology, I will call the above conjunction of ideas 1 and 2, in combination with the libertarian conception of control, the libertarian conception of freedom, or libertarian freedom for short.

But this is a suspect idea.

Is the libertarian conception of freedom coherent?

Many philosophers hold that this conception of freedom is incoherent.

Many people think, like Leucippus did, that

> naught happens for nothing, but everything from a ground and of necessity[1]

When asked to expand on this claim, these people tend to claim something like that for every occurrence there must be an explanation, and the explanation invokes earlier occurrences that necessitated that the event that is being explained happened, and not

[1] In another translation: "Nothing occurs at random, but everything for a reason and by necessity." (Kirk, Raven, Schofield **569**, 1983, p. 420; Fr. 2. Aetius I, 25, 4). They say it is "the only extant saying of Leucippus himself". The translation in the main text is from Russell's *History of Western Philosophy* (1972).

something else, and that it happened at that particular time and not some other time.

Whether this view of the causal ordering of events should apply also to the events that happen inside our souls, has always been a point where intuitions diverged.[2]

But it is unclear how much is gained, in terms of freedom, if the principle of universal causation is fended off from the mental realm. In the last century quantum mechanics has shaken the view that it applies at all, even to the physical realm. On the orthodox quantum mechanical view of the causal ordering of the world physical events do happen without a necessitating cause.[3] However, it is quite natural to assume that, to the extent they are underdetermined by their physical causes, quantum events are random. This is what we commonsensically seem to mean by randomness: objective underdeterminedness. Now the principle that every occurrence is either determined or, to the extent it is underdetermined, random, seems to apply irrespective of whether we are considering physical events or events belonging to a genuinely non-physical mental realm. It strikes us as a simple conceptual truth.

A random occurrence is a paradigm case of uncontrolledness.

Some philosophers take it for granted that determinedness and randomness exhaust the field of options. Kant, for example wrote in *The Critique of Practical Reason* that

> [I]f one attributes freedom to a being whose existence is determined in time, it cannot be excepted from the law of natural necessity of all events in its existence, including also its actions. Making such an exception would be equivalent to delivering this being to blind chance. Since the law inevitably concerns all causality of things so far as their existence is determinable in time, freedom would have to be rejected as a void and impossible concept if this were the way in which we thought of the existence of these things as they are in themselves[4]

---

[2] Socrates in the *Phaedo* seems to consider the phenomenon of self-determination as a ground for arguing for body-soul dualism.

[3] But note that not all interpretations of quantum mechanics are indeterministic. On this issue see the Appendix.

[4] Kant 1788/1949, p. 201. (Part I, Book 1, Chapter 3.)

Indeed, the principle that an event is either of necessity or of "blind chance" can be said to be part of common sense. To prove this principle wrong some underdetermined events, our supposedly libertarian free choices about how to set our will, should be distinguished somehow from random events. The libertarian is required to explain how these events are non-random, although they are underdetermined. This explanation should be instructive about how activity is achieved in these cases of setting our will, in contrast to the passivity of determinedness, and the passivity of randomness, and about how these events are controlled by us. The distinction between determined and random events, on the one hand, and events of this third kind, on the other, should be relevant for the values that are normally associated with freedom, such as self-determination, and moral and intellectual responsibility.

Can a libertarian free choice be rational?

Some philosophers say that, even if the coherence of the libertarian conception of freedom is granted, this is not a very valuable kind of freedom because we can be free in the libertarian sense only insofar as we are nonrational.

Let us take a toy model of rational control.

We have reasons prior to a choice in the light of which our choice is either rational or not. The libertarian does not want a determining relation between the reasons and the choice, because his theory of freedom requires the freedom from determination by all previously existing things. So the libertarian wants the choice to be free, to some extent, from the reasons. Or, alternatively, he may allow for the determination of the choice by the reasons, but then he must insist that some of the reasons should have been freely chosen previously.

Let us consider first the latter possibility. What would it take for a reason to be chosen freely in the libertarian sense? The worry is that adopting a reason freely is adopting a reason for no reason at all, so our choosing of the reason to be a reason on which later we perhaps act is nonrational. How could it be rational? It could be rational if it was the case that we adopted the reason on reasons we already had. But then the same question would arise in respect of these reasons.

The attempt to secure both libertarian freedom and rationality at the same time this way leads to an infinite regress.[5]

Let us see now the other possibility. This is that freedom intervenes somehow between the reasons and the choice. Indeed, it is plausible that the reasons we have (desires and beliefs about how desires can be realized in the given circumstances) do not control us mechanically. We consider some of them, and neglect others. Among those we consider we may give more weight to some than we give to others. In our toy model we can view it as having the capacity to adopt a function that associates weights to reasons (a weight can be zero in the case when the reason is not considered), assuming that we adopt this function freely in the libertarian sense. Now the worry is the same as in the previous case: that we adopt this function nonrationally. If we want the adoption of the function to be rational, and also free in the libertarian sense, we embark on a regress similar to that in the previous case.

Either we adopt some reasons freely in the libertarian sense, or adopt a function freely to weigh reasons, we are rational in the sense that we have reasons that rationalize our choice. We are not fully rational, though, since not every mental operation we perform is answerable to reason. We are not limitlessly free either, since we work with a raw material that is given: the reasons that we have passively, or, if some of them were adopted freely, the reasons that suggested themselves for adoption. We do not construct ourselves *ex nihilo*. Neither is a problem in itself. But maybe this is a problem: although we can be said to be both rational and free (in the libertarian sense), we are free only to the extent we are nonrational, or, conversely, we are rational only to the extent we are unfree. The worry is that then what the power to be free in the libertarian sense comes to is

> some entirely non-rational (reasons-independent) flip-flop of the soul,

as Galen Strawson put it.[6]

The causal conception of control and freedom

---

[5] Galen Strawson 1986.

[6] Ibid, p. 54. "The agent-self with its putative, freedom-creating power of *partially* reason-independent decision becomes some entirely non-rational (reasons-independent) flip-flop of the soul."

The alternative to the libertarian conception of control is the idea that control is a special case of the causal determination of what is to come by what is already laid down.[7] This special case obtains when what is to come is determined by causes which belong to a special kind, and by some virtues characteristic of this kind, say, because with these causes (which are parts of our mental life, maybe operating as a conscious mechanism of a specific form) we can identify or be reasonably identified[8], determination by them is not worrisome for freedom, quite to the contrary, it promotes freedom.[9] I call this idea the causal conception of control. I will call the idea of freedom involving the causal conception of control the causal conception of freedom.

It is uncontroversial that this idea can be applied to the freedom of action, in which case the special kind of causal determination that promotes freedom is determination by a mental state that we may call a will.

But now the question is whether the same conception of freedom can apply to the setting of the will itself.

The shallowness of the causal conception of control and freedom

Applying the causal conception of freedom to the will would mean to say that the will is determined, not by just any cause, but, again, by something in us (see footnote 8), whose determining role

---

[7] When I talk about the determination of what is to come by what is already laid down, what I have in mind is what is often called "event causation". The idea known as "agent causation" is of course a variant of the libertarian conception of control, when the will gets set without being determined by what is already laid down.

[8] The exact quality of causes which I referred to simply by saying that with them "we can identify or be reasonably identified" varies from one causal theory to the other. It can simply mean that control is the determination of the will by causes which are our truly embraced desires (desires we desire or at least agree to have), or beliefs about who we are or want to be. Or by causes that constitute our character (as it is in Hume's account of freedom). Or by a specific mental faculty whose designated function is control (the deliberative faculty). Or by a reason-responsive mental mechanism for which we had "taken responsibility", meaning that we previously accepted, at least tacitly, that if our will is determined by it then we are responsible for what we will and what we do on our will (Fischer and Ravizza 1998).

[9] As one of the proponents of the causal theory, Alfred Ayer put it: [I]t is not when my action has any cause at all, but only when it has a special sort of cause, that it is reckoned not to be free (Ayer 1954).

does not breach but promotes freedom. This we may call a second-order will.

However, by positing this criterion for the freedom of the will, we launch ourselves on a regress, unless we are ready to accept a certain shallowness of control and freedom. Our notion of the freedom of the will would be shallow if we did not require also the freedom of the second-order will by which it is determined. But as a criterion for the freedom of the second-order will, if we don't want to bring in the libertarian conception of freedom at this point, we have to posit determinatedness by a third-order will. And so on. If we quit this regress at any point (and not by accepting the libertarian conception of freedom), we agree that our action is the necessary causal consequence of causes with which we cannot reasonably identify or be identified. So it is a passion after all. If we do not quit the regress we will end up with the conclusion that every free action is preceded by an infinite series of willings (the obtaining of mental states or processes with which we identify or can be identified). But this is absurd.[10]

It should be noted that nothing specific about the nature of the "will" has been exploited in this regress argument. If control is analysed as determinatedness by the "right kind of thing", whatever the exact criteria for a thing to be of the right kind would be (it can even change from order to order[11]), after a finite number of steps backwards in the causal chain it turns out that we have to fall back to either of two kinds of uncontrolledness: a) when that thing "of the right kind" occurs ultimately at random and b) when that thing is ultimately determined by causes "of the wrong kind". If we are

---

[10] "[I]f my volition to pull the trigger is voluntary, in the sense assumed by the theory, then it must issue from a prior volition and that from another *ad infinitum*." Gilbert Ryle wrote this in Chapter 3, Section 2 of *The Concept of Mind* (1949, p. 67). The credit for the argument is sometimes given to him altogether, so, for the sake of simplicity, I will sometimes refer to it as "the Rylean regress argument". Yet, I think, that conceiving of freedom as determinatedness by the will leads to a regress if this conception is applied to the freedom of the will is so obvious that it must have been pointed out very early in the course of Western philosophy. Thomas Reid discussed this argument as a possible (and answerable) objection to his theory of free will (1895, p. 501). Thomas Hobbes also thought it was straightforwardly absurd to talk of the freedom of the will because he construed freedom as determinatedness by the will. He wrote: "I acknowledge this liberty, that I can do if I will, but to say, I can will if I will, I take to be an absurd speech" ("Of Liberty and Necessity", 1654/1962).

[11] It changes from order to order in a classic discussion of the impossibility of an infinite regress of causes which are all "of the right kind" by Alfred Ayer in his "Freedom and Necessity" (1954).

reluctant to accept this shallowness of control and freedom, our dissatisfaction with it launches us on an infinite regress.

The conclusion to be drawn from the regress problem is that on the causal conception of control there is a limit to the extent or depth to which control penetrates the process of will-formation. The proponent of the causal conception must hold that the causal mechanism that led to our will's being set is irrelevant after a few steps taken backwards along the causal chain to the question whether our will is free or not.

Is the shallowness of control a problem?

Kant, for example, was strongly dissatisfied with this feature of the causal conception of freedom (which for him was exemplified by Hume's theory), and thought that if freedom was so conceived,

> it would in essence be no better than the freedom of a turnspit, which when once wound up also carries out its motions of itself.[12]

Some proponents of the causal conception of control, however, argue that this shallowness of control is not a problem. The process of will formation serves practical ends, so it should not take too long. Surely, it must be a finite process. The control mechanism is presumably a reflective-evaluative process in which we reflect on mental items (moods, inclinations, character traits, desires, higher-order desires, judgements about the desirability of some ends, beliefs about the possible means to those ends given the way we believe the world is, etc.) that, in our subjective perspective, seem to have the potential to move us to act in certain ways if we let them. It is unreasonable to require that we should reflect on and evaluate the whole of our relevant psychology. Some parts of our motivational structure should be given and operate unreflectedly for our deliberations to be practical.[13]

Some libertarians, on the other hand, argue that this shallowness of the causal conception of control and freedom results in the vulnerability of agents who are free on the causal account to "covert

---

[12] Kant 1788/1949, p. 203.
[13] Dennett 1984a.

non-constraining manipulation". This is how Robert Kane describes the situation:

> [In cases of one agent being controlled by another the non-constraining way] the controllers do not get their way by constraining or coercing others against their wills, but rather by manipulating the wills of others so that the others (willingly) do what the controllers desire. The controlled agents consequently do not feel frustrated or thwarted. They act in accordance with their own wants, desires or intentions. … In the most interesting cases, such control is a "covert" nonconstraining control…in which the controlled agents are unaware of being manipulated or perhaps even unaware of the existence of their controllers.[14]

The worry is that, intuitively, agents so manipulated are unfree, whereas, due to the shallowness of the causal conception of control, their being so manipulated is consistent with their being free on the causal conception of freedom.

Either control or alternatives

Another important consequence of construing control as a kind of causal determination is that ideas 1 and 2 can never apply to the same event. No event of our mental life can be both controlled and such that there was a possible alternative to it given everything that took place previously. On this conception alternatives can be bought only at the price of forfeiting control.

Some adherents of the causal conception, however, made attempts to show that this is not the case. They argued that it is possible to have alternatives even if all events are determined by previous facts and events, in short, even if the world is deterministic.[15]

If it was true, then the causal conception of freedom would be a very attractive option, perhaps a clear winner: it would mean having both alternatives and control without the difficulties of making sense of non-causal control.

---

[14] Kane 1996, pp. 64-5.

[15] Cf. Widerker 1987, Vihvelin 1988, Kapitan 2002, J. T. Saunders 1968, Fischer 1984, 1988, Lewis 1981.

If it isn't true, however, then it is not only the case that the causal conception of control is incompatible with there having been genuine alternatives to the event whose coming about is controlled by us, and that so we have to abandon idea 1, but also that control is not the kind of control specified by idea 2. According to idea 2 control is controlling which of the alternatives should occur. Adopting John Martin Fischer's terminology[16], control in this sense I will call regulative control. If control is conceived as a species of causal determination, and if there aren't really possible alternatives to what is causally determined, then control so conceived cannot be regulative. Control could mean only that the agent (causes "within" the agent with which the agent can be reasonably identified, or with which the agent does identify, or with which the relevant moral community identifies the agent) plays a necessary role in the causal production of the will (and so of the action). Control in this sense, following Fischer, will be called guidance control.

Do we need alternatives?

Other proponents of the causal conception of control admit that construing control as a species of causal determination amounts to giving up on genuine (i.e. objectively possible) alternatives, but argue that it does not really matter.[17]

One way of arguing that it does not matter is to argue that the values that we associate with freedom, most importantly moral responsibility, contrary to philosophically unguided intuition, do not really require the existence of genuine alternatives.[18]

One variant of this position is to claim that freedom requires that there be alternatives only in the "hypothetical" sense. That is to say that

> If we choose to remain at rest, we may; if we choose to move, we also may. Now this hypothetical liberty is universally allowed to belong to everyone who is not a prisoner and in chains.[19]

---

[16] Cf. Fischer 1995, or Fischer and Ravizza 1998.

[17] Very expressive is the title of one of Dennett's articles advocating this view: "I Could Not Have Done Otherwise—So What?" (1984b).

[18] Frankfurt 1969, Dennett 1984a, 2003.

[19] Hume 1975, p. 104.

This is essentially the requirement that actions covary with the will.

Some developed this intuition to the theory that freedom and moral responsibility do have a could-have-done-otherwise condition, but the sense of "could have done otherwise" that is relevant for freedom and responsibility is captured by an analysis of it in terms of a counterfactual conditional, according to which to say that I could have done otherwise is to say that I would have done otherwise had I so willed.[20]

Another variant is to claim that we should keep to the Kantian distinction between the practical and the theoretical perspectives, and that practical possibility, i.e. the existence of alternatives in the practical but not in the theoretical perspective, is enough for freedom:

> A person is free if she is capable of determining her actions through practical reasoning,[21]

and in practical reasoning

> we cannot possibly employ a conception of the alternatives that are available to us that is narrower than the set of actions that *we would perform were we to choose to do so*.[22]

Freedom in the sense of choosing from alternatives is a correlatum of a specific point of view also on a theory according to which it is epistemically, scientifically and practically relevant to adopt a variety of stances when one describes, understands and predicts a system. There are systems towards which it is sometimes useful to adopt a personal stance, in which we view them as free agents contemplating and evaluating different courses of actions, and selecting from them. When adopting this stance we treat these systems, among other things, as representing their environment, as having goals or values, as having alternatives, and as being morally

---

[20] Moore 1912.

[21] Bok 1998, p. 120. An important difference though between the views of Kant and modern perspectivalist compatibilists is that Kant thought that the practical perspective corresponds to how things really are (*noumenal* reality), and the theoretical perspective corresponds to how they appear in experience (*phenomenal* reality).

[22] Ibid, p. 117, stress mine.

responsible. The legitimacy of adopting a certain stance comes solely from the given stance's pragmatic value in understanding and predicting the system and interacting with it. Adopting the personal stance towards certain systems such as people has a clear pragmatic value. In many cases it is practically impossible to describe and predict them relevantly in the physical stance, in which they are described as mechanisms. There is nothing more to freedom and responsibility than the usefulness of the personal stance.[23]

There are philosophers who argue that alternatives in the objective sense would not make us any freer than having alternatives in one of the above senses.

The question is whether these kinds of freedom that have alternatives, at best, only in some perspective which is not the objective "physical" or "theoretical" perspective, can ground the values that we associate with freedom, such as self-determination, moral responsibility and intellectual responsibility, or whether they ground them equally well as conceptions of freedom that have genuine alternatives can.

Can we have alternatives, anyway?

It is an argument in favour of the causal view if it is true that we could not have genuine alternatives anyway, for reasons related to the metaphysics of our world. If so, than ideas 1 and 2, of which I claimed our commonsense conception of freedom to consist in, both come to nothing, and alternatives in the hypothetical (subjective, practical, stance-dependent) sense and control in the sense of guidance control is the most we can hope for.

Arguments related to time

It is reasonable to say that the existence of genuine alternatives would require a part of the future to be up to us, therefore it would require a part of the future to be objectively contingent. Some philosophers say that no part of the future is objectively contingent for reasons related to either the concept or the empirical nature of time (or spacetime).

---

[23] Dennett 1984a, 1987.

Some of the arguments for this claim are ancient like the one discussed by Aristotle in the ninth chapter of *De interpretatione*, often referred to as the logical fatalist argument, which has the law of the excluded middle among its premises, together with two principles about time: that the law of the excluded middle applies also to propositions about the future, and that the past is ontologically fixed and so unchangeable.[24]

There is another very ancient tradition of thought that can be traced back to Parmenides of Elea that denies the existence of future contingencies by arguing that everything, past, present or future, exists on a par; for the future to be open it should be ontologically different from the past, which is fixed; but such a difference could be real only if becoming was real, and becoming is an incoherent idea. This view had a famous modern adherent in the person of Ellis McTaggart.[25]

I don't think either of these arguments pose any serious threat to future contingencies and so to libertarian freedom. There is, however, a much more promising line of argument to the Parmenidean conclusion from the special theory of relativity.

Time has to be an A series, in McTaggart's parlance, in order to flow and bring along events which are not yet. A-determinations have to have an objective meaning to correspond to different ontological determinations.

But A-determinations are observer-relative according to the special theory of relativity. The notions of the future and the past are dependent on the notion of the present, the alleged locus of becoming, the transition between the ontological determinations characteristic of the future and the past. The notion of the present, in turn, is dependent on the notion of simultaneity. But simultaneity has a meaning only relative to a frame of reference. If you and I are inertial observers here and now in relative motion, then your present – the set of events that are simultaneous with our meeting here and now in your frame of reference – consists of events that are different from those that make up my present. Our presents are three-dimensional planes (subspaces) in four-dimensional spacetime, and mine is tilted relative to yours. There are events which are past

---

[24] The credit for first presenting the fatalist argument as an argument from the unchangeability of the past is traditionally given to Diodorus Cronus. (See Epictetus, *Dissertationes* II 19, 1-5 in Döring 1972.)

[25] McTaggart 1908.

relative to your present and future relative to mine. There is no principled reason to prefer your frame to mine or mine to yours. But then there is no objective lapse of time and A-determinations cannot correspond to anything ontological.

So the worry is that relativity theory must be false for alternatives to be real and libertarian freedom to be possible.

The pioneers of relativity theory were aware of the bearing of the relativity of simultaneity on the metaphysics of time.[26] Cassirer and Gödel devised arguments from it to prove that time had no *noumenal* reality in Kant's sense.[27] Yet, the discussion of the question heated up only about forty years ago by two influential papers of Hilary Putnam and Wim Rietdijk.[28] In response to these arguments different ways have been suggested either to resist the simultaneity of relativity despite of the body of empirical evidence to which the special theory appeals, or to reconcile the relativity of simultaneity with objective becoming. A consensus is nowhere near.

The view that determinism is a (near) scientific fact

Many philosophers treat determinism as an empirical fact, or a hypothesis that has been so strongly corroborated that no theory of freedom that is incompatible with it is worth much attention.

It could be a fact of empirical psychology, as Hume claimed it was. It could be the case that every action is determined by an intention or will to act, every intention or will is determined by a prior psychological-motivational state, and every motivational state is determined ultimately by factors in respect of which we are passive: genetics, upbringing and environmental stimuli, that shape our psychology.

It could also be a fact of physics. Most actions involve bodily movements which are physical events. Surely, libertarians want freedom in the sense of having genuine alternatives in respect of

---

[26] See for example Carnap's note on his discussions with Einstein about the question (Carnap 1963, p. 37) to be quoted later on p. 158.

[27] Cassirer 1920, Gödel 1949a. Their Kantian interpretation of relativity theory is reviewed by Mauro Dorato 2002.

[28] Putnam 1967, Rietdijk 1966. Rietdijk's article in the *Philosophy of Science* was titled "A Rigorous Proof of Determinism Derived from the Special Theory of Relativity", indicating a direct bearing of the problem on our subject. (Although the term 'determinism' is used by Rietdijk in a sense that is different from the normal usage. Nothing like causal or nomological determinism can be derived from STR.)

these, too. If determinism holds universally in physics, these events can be traced back causally to the Big Bang. From the determinism of physics its dynamical closure follows. If so, all mental events that have a role in the causal production of action must supervene on physical events, and so their evolution must be deterministic, too. For nearly three hundred years the very successful mathematical methodology of classical physics supported the determinist hypothesis strongly. With the rise of quantum mechanics the question now seems to be open.

But determinism can be a fact of neuroscience, too. If the evolution of mental states supervenes on the evolution of neural states, then the determinism of the latter ensures that of the former. There are philosophers who think that it is a scientific fact that the evolution of neural states is deterministic and it makes the question whether physics is deterministic irrelevant.[29]

Compatibility with determinism as a norm of theory choice with respect to freedom

For some philosophers the mere *possibility* that either physics or our neuroscientific description is deterministic is enough reason to hold that compatibility with determinism is a norm to which our thinking about freedom should be adjusted.

John Martin Fischer, for one, in the beginning of *The Metaphysics of Free Will*, points out that what is at stake, ultimately, in our thinking about freedom, is the meaning of our personhood.

He then invites us to consider the possibility that the consortium of top scientists announces that the world is deterministic. Fischer's intuition is that we would not react to such an announcement with abandoning fundamental attitudes toward ourselves as person. For example, we would not give up the practice of ascribing moral responsibility. This is not just a psychological and practical social incapability[30]:

> Rather, I am making a *normative* point. I am saying that, upon reflection, it just does not seem appropriate or plausible to think that we should abandon our view of

[29] Honderich 2002.

[30] As an earlier adherent of a similar view, Peter Strawson argued (1962).

> ourselves as persons, if it turned out that the consortium of scientists were correct.[31]

It seems to me that Fischer derives a norm of theory choice from this insight. Our view of ourselves as persons and moral agents is very firm. Meanwhile, it is possible, for all we presently know, that the world is deterministic. So we should develop a metaphysics of freedom and moral responsibility that secures it that freedom and anything that depends on freedom be compatible with determinism.

My position and the plan of the thesis

I am no friend of philosophical policies like the one recommended by Fischer. I think we should aim at a metaphysical theory of freedom which captures best what we naturally think freedom consists in. If it is found out that such kind of freedom we cannot have, then we should face the loss. As philosophers we should aim at the truth, and not at a theory that is bearable.

If there is a plurality of the senses in which we naturally take ourselves to be free, then we should account for the values central to our personhood in a differentiated way. We should sort out which of them is tenable on which conception of freedom, and if some of these conceptions of freedom are found impossible in our world, then we should face the loss of the values that are associated exclusively with these notions of freedom.

So I think the questions that were raised in relation to alternatives and control in this introduction need to be answered. These were *a)* whether we can have genuine alternatives if control is achieved by

---

[31] Fischer 1995, p. 7, stress in the original. A similar consideration is reiterated in Section II.2 (pp. 253-4) of the concluding chapter of *Responsibility and Control: A Theory of Moral Responsibility*, a book Fischer published together with Mark Ravizza in 1998, with an important modification. There they are not making the normative point that we *should not* react to the announcement of determinism by top physicists with abandoning the view of ourselves as moral agents, implying a norm of theory choice for a theory of responsibility. They only record it, among other advantages of their account of responsibility that it is compatible with both determinism and its falsity, so they can look forward to the physicists' discovering the truth about determinism confidently. So modified, without the normative edge, this consideration is much more acceptable. It is really a valuable feature of a theory if its consequences are insensitive to how an open empirical matter will eventually be settled. However, this feature in itself doesn't make a theory more likely to be a good approximation of the truth than a rival theory lacking this feature.

way of causal determination of what is to come by what is already laid down; *b)* whether the shallowness of the causal conception of control that has been pointed out is a problem for self-determination; *c)* whether the values we associate with freedom, such as moral and intellectual responsibility, require genuine (objectively possible) alternatives; *d)* whether we can have genuine alternatives in this world, for all that our best scientific theories have to say about this matter; *e)* whether the libertarian conception of control is a coherent one; and *f)* whether control, on the libertarian conception of it, can be rational.

In chapter 2 I will argue, in response to question *a*, that *if we adopt the causal conception of control we do forfeit genuine alternatives*. This conclusion seems to follow form a very obvious consideration: It is not in our power to change the past. If control means that some past facts and events determine how we shall choose (by the force of some causal laws), we cannot help that either. So it is not in our power to do any other than we actually do. I will argue, against the ingenuity of "local miracle" and "multiple pasts" compatibilists, and those who doubt that our inability to change the past or the causal laws transfers to what the past and the laws taken together entail, that this simple commonsense argument stands the test of philosophical scrutiny.

In chapter 3 I will argue, in response to question *b*, that *the shallowness of control, on the causal conception of it, is a serious concern for self-determination*. I will argue, against Dennett, that this shallowness is not a necessary condition for self-determination to be practical, and that the fact that this shallowness makes freedom on the causal conception of it is coherent with what Kane called "covert non-constraining manipulation" is a problem.

In chapter 4, in response to question *c*, I will argue that *the senses of moral responsibility that are available without genuine alternatives leave our intuitions about what moral responsibility requires unsatisfied*. My main opponent in this chapter will be Dennett again. In his two books about freedom *Elbow Room* and *Freedom Evolves* he offered arguments to the effect that deterministic agents can be correctly held responsible. I will show that his arguments fail to give strong enough support to this claim. Then I will argue, drawing on van Inwagen, and against Fischer's objections, that the famous "Frankfurt-type cases" are not really counterexamples to the "could-have-done-otherwise condition" of moral responsibility. I will refer to Martha Klein's work showing that our moral intuitions operating in practical cases do

support a could-have-done-otherwise (or at least an "ultimate responsibility") condition for culpability. In chapter 8 I will add to these arguments that even those compatibilists who claim that it is fair to hold people accountable for whatever proceeds from their character in a necessary way should accept the principle that no one is more responsible for a response one's character produces necessarily to a stimulus than one is responsible for one's character, provided that one is not responsible for the stimulus. I will argue that once this principle is accepted there is no way to avoid the principle of the transfer of non-responsibility through causal necessitation.

In chapter 5, again in response to question *c*, I will argue that *we cannot properly be said to possess rationality and intellectual responsibility unless we are free in the libertarian sense*. My argument will be based on an ancient argument by Epicurus, whose intuition was that if there is a causal explanation for why a thought came about, then a rational explanation for the same cannot be simultaneously true. Against a modern version of the argument given by C. S. Lewis, Elizabeth Anscombe objected that the causal and rational explanations are independent matters that do not compete with each other. Alternatively, the Epicurean argument can also be attacked the Davidsonian way, i.e. by claiming the reasons are causes, and so the causal and the rational explanation account for the same genealogy of thoughts. I will try to show in chapter 5 that both the Anscombian and the Davidsonian objection fail.

In response to question *d*, in chapter 6 I will argue that, *as far as our present knowledge goes, determinism is empirically unfounded*. I will argue that if there is evidence for the determinism of our mental life, contrary to what Hume famously claimed, this evidence must come from "below", i.e. from reductionism and the determinism of "the underlying reality". I will argue against physical determinists that there is no empirical evidence that physical systems containing, or interfering with, conscious minds would evolve deterministically. Against neural determinists of Honderich's fashion I will argue that there are conceivable ways quantum indeterminacies could propagate to the macro level of brain states, and that the objection by Papineau and others to the theory that these quantum indeterminacies may make room for the mind to freely control its brain would contradict quantum mechanics because it would ruin the Born rule rests on a mistake related to the nature of probability.

Also in response to question *d*, in chapter 7 I will argue that *as far as our present knowledge goes, the future may well be open*. As I have already indicated, my main worry concerning the openness of the future is the argument from the special relativistic invariance of inertial observers as offered by Gödel, Putnam and others. In chapter 7 I will present ways the relativity of simultaneity can be resisted consistently with the body of empirical data that led to relativity theory, or, alternatively, how the relativity of simultaneity can be lived with, consistently with the idea of objective becoming. My preferred solution to the puzzle posed by relativity theory is to adopt a relativistically invariant local notion of the present, and to conceive the past and the future on the basis of the lightcone structure, which is also invariant. I will defend this theory advocated by Howard Stein and Dennis Dieks against Simon Saunders's objections that it will fail to meet a relativistically meaningful requirement of intersubjectivity concerning determinations that are meant to have an ontological significance, and that so it will lead to solipsism.

In chapter 8, in response to question *e*, I will argue that *the libertarian conception of control involves no contradiction*. I will argue though that libertarianism makes sense only if a radically non-reductionist ontology of persons is adopted. I will review the alternative, ontologically less extravagant libertarian attempts to save a class of undetermined events from randomness, most prominently Kane's and Nozick's causal indeterminism, and will conclude that they fail to distinguish between freedom and randomness in a morally relevant way. I will also argue that David Wiggins's suggestion that some undetermined choices are non-random *because* they are intelligible in the light of the agent's reasons, cannot be the solution to the coherence problem of libertarianism either.

In response to question *f*, also in chapter 8, drawing on the results of chapter 5, I will argue that *control conceived the libertarian way can be rational, in fact, only libertarian control can be rational.*

So my thesis will be both a positive argument for the libertarian conception of freedom on the basis that it grounds personhood in a much fuller sense than the causal conception of freedom can, and a defence of it against the objections that it would be incoherent, or necessarily non-rational, or empirically impossible.

There is a significant philosophical price though to pay for libertarian freedom. In the concluding, ninth chapter I will assess how

great this price is, and try to compare it to the price of the causal (compatibilist) theory.

2 Can We Have Genuine Alternatives If Control Is Achieved Causally?

The consequence argument

Some philosophers argued that it may simultaneously be true that a) there are genuine alternatives to how we set our will, and b) that we control how we set our will, even on the causal conception of control, i.e. even if controlling our will is achieved by way of its causal determination by causes of some "right kind". The philosopher who wants to show this cannot appeal to the possibility, if there is such a possibility, that those causes could be some other way than they actually are. That is irrelevant. The question is whether we have genuine alternatives given what we are and how we are at the time of setting our will. So the philosopher who claims to be able to show that this is possible must be ready for the task of showing that we can have alternatives even if the world is deterministic.

It is important now to clarify what exactly we mean by determinism.

In the sixth chapter I will argue that there is no direct evidence on the psychological level that our intentions or wills to act would be determined by prior mental states and environmental circumstances.

If the determinism of the mind cannot be established empirically on the ground of facts observed at the psychological level, then the evidence for determinism, if there is one, must come from below, that is, from the study of a "deeper" ontological level that is found deterministic, and to which the psychological level is thought to reduce. There are two candidates for the role of such an underlying ontological level: the level of the neurons and the level of fundamental physical entities.

I propose to use the following definition of physical determinism:

The set of all physical events (U), of which the set of actions is a subset (A), has the property that there are core subsets within U, such that with a core subset and with the laws of physics only one totality of U is logically coherent, therefore, only one subset A is coherent; and the set of all events which are past relative to any arbitrarily chosen moment of time in any arbitrarily chosen frame of reference is such a core subset.

The analogous definition of neural determinism is:

The set of all neural events of an agent (N), of which the set of (the initiation of) all his actions is a subset (A), has the property that there are core subsets within N, such that with a core subset, with the laws of neuroscience, with a given input from the sense organs, and with a given totality of other physical influences coming from outside the neural system[32]*, the latter two treated as a fixed set of boundary conditions, only one totality of N is logically coherent, therefore, only one subset A is coherent; and the set of all events which are past relative to any arbitrarily chosen moment of time is such a core subset.*

It is controversial if these definitions capture what is usually called *causal determinism*. They do if the nomological account of causation is assumed. But it is not necessary here to argue for the nomological account of causation. Maybe causation is something deeper than what is captured by the nomological account. The scientific evidence, however, to which determinists refer is on this "superficial" level. Those who believe in determinism believe in it because the scientific prediction or retrodiction of events on the basis of known events and natural laws has been a success in many areas of research. These formulations of the determinist hypothesis are generalizations of the success that science has achieved in predicting and retrodicting the evolution of a large number of systems that it studied. The determinist hypothesis is a possible explanation for this success. I know of no other empirical evidence of "causal" determinism that would stem from the view that causation is a deeper reality than the subsumption of types of events under general laws. So I propose to use the adjectives "causal" and "nomological" interchangeably as long as determinism is considered as a hypothesis for which we expect scientific justification.

From the determinism of the underlying realm so understood the reducibility of the part of the mental realm that plays a role in action production to that realm follows, as long as it is secured that an action has a description as an event of the underlying realm. For if this definition of determinism is accepted, the causal closure of the

---

[32] It is possible that "the neural automaton" is influenced by other physical factors apart from the input coming from the senses. Chemicals in the blood, metabolic disorders, radiation, heat or cold, mechanical disturbances, etc. can all affect the neural system. However, they don't seem to be relevant to the question of freedom. If mental states supervene on neural states, and the evolution of the neural states is deterministic (given the sensory input), apart form such external physical influences (which we may allow to be indeterministic), then these influences do not make us freer than we would otherwise be. So it seems useful to define the determinism of the neural system relative to these, treated as boundary conditions.

underlying realm in question follows from its determinism. If the physical realm is deterministic under the above definition, then it allows for no interference from outside to make a difference to its evolution. If our neural system is deterministic on the above definition then it is causally closed apart from the input from sense organs and other outside physical influences, and so it does not allow for the irreducibly mental to interfere in its evolution. If the causal closure of an underlying realm that is evidently involved in actions is established, then the part of the mental realm that is involved in action production must reduce to it, otherwise it could not have any role in bringing about actions. It is uncontroversial that if we want alternatives, it means that we want alternatives to willings that give rise to courses of actions that have neural or physical aspects. So we don't need an independent argument for reductionism if we have evidence for the determinism of either the neural or the physical realm on the above definition.

It seems evident that genuine alternatives are impossible if determinism is true. It hardly needs an argument, if determinism is understood as it was suggested above.

Yet, the argument has been given formally and has been discussed at great length by a good number of philosophers. When it is stated formally, the argument is usually called the consequence argument.[33]

The first two premises of the argument look to be platitudes:

1 We are unable to make the past different from what it is.

2 We are unable to make the laws of nature different from what they are.

The third premise is determinism:

3 Past events are *core* in the set of all events, past, present and future, in the sense specified above: with them and with the laws of nature only one totality of events is possible.

From the conjunction of these three premises it follows that

[33] To my knowledge the first philosopher to give this formal argument was Kant in *The Critique of Practical Reason.* The most frequently referenced proponent of the argument in present day discussions is perhaps Peter van Inwagen (1983).

4 We are unable to make the present or the future different from what it is, which is what the past and the laws of nature prescribe.

Therefore

5 We could not do any different from the way we actually do.

Philosophers who are compatibilists about alternatives and determinism challenged every component of this argument that can be challenged: the logical entailment from 1, 2 and 3 to 4, premises 1 and 2, and the logical entailment from 4 to 5.

Attacking the logical entailment from 1, 2 and 3 to 4: the "transfer principle"

Attacking the entailment from 1, 2, and 3 to 4 is denying that the modal operator *"unable to make … different"*, which henceforth I will call the "inability operator", is closed under conjunction introduction and logical entailment. Some critics of the argument say that the closure of the inability operator is a hidden premise of the argument, and that it is false.

Let $N_{S,t}(p)$ denote the modal operator that *"person S at time t cannot make p different"*.[34] The closure of $N_{S,t}$ under conjunction introduction and logical entailment enables us to draw from premises $N_{S,t}(p)$ and $N_{S,t}(p \supset q)$ the conclusion that $N_{S,t}(q)$, for any two propositions p and q.[35] (In our case p would be the conjunction of the laws of nature and the initial conditions—the past, q would be the present and the future, containing an action of the agent.) Although the closure principle is intuitively very appealing, David Widerker managed to

[34] I follow the notations of O'Connor 2000 (which are the same as in Fischer 1995).

[35] It can be easily proven:

1 $N_{S,t}(p)$ (Premise. In our special case p is the conjunction of the past and the laws of nature. But in the present deduction nothing depends on that.)

2 $N_{S,t}(p \supset q)$ (Premise. In our special case $p \supset q$ is the determinist hypothesis. But in the present deduction nothing depends on that either)

3 (Therefore:) $N_{S,t}(p \ \& \ p \supset q)$ (From 1, 2, and closure under conjunction introduction.)

4 $\Box ((p \ \& \ p \supset q) \supset q)$ (This is a logical truth.)

5 (Therefore:) $N_{S,t}(q)$ (From 3, 4, and closure under logical entailment.)

show some counterexamples to it.[36] Here is one of them in Timothy O'Connor's presentation[37]:

> Suppose that by destroying a bit of radium at t_1, Sam prevents its indeterministically emitting a subatomic particle at t_2. Suppose further that this is the only way by which Sam can make sure that it won't emit radiation at t_2.
>
> If we let p = The bit of radium does not emit a subatomic particle at t_2, and q = Sam destroys the radium at t_1, then [the inference principle that $N_{S,t}(p)$ and $N_{S,t}(p \supset q)$ entails $N_{S,t}(q)$] licences us to conclude that Sam was unable at t_1 to refrain from destroying the radium, for both the needed premises are satisfied. Clearly, Sam did not have control over the truth of p—he couldn't ensure that a particle was emitted at t_2, even thought this might have occurred had he not destroyed the radium. So $N_{Sam,t1}(p)$. Consider now the second premise, $N_{Sam,t1}$(if p, then q). This also holds because the conditional (if p, then q) is true and its truth was not within Sam's control. To have control over the truth of the conditional, Sam must have been able to make it the case that *not* (if p, then q). This is equivalent to (p and not q). If Sam had made true the second conjunct (not-q)—that is, had he refrained from destroying the radium—then he would have no means of ensuring that the first conjunct (p) is also true (though...this might have been the case nonetheless). But surely, it is consistent with these facts about the example to suppose...that Sam was able to falsify q, that is, to not destroy the radium at t_1. Therefore, as stated, the inference rule [the principle that $N_{S,t}(p)$ and $N_{S,t}(p \supset q)$ entails $N_{S,t}(q)$] is invalid.

Now, since the inference rule is the logical consequence of the closure principle, the counterexample to the inference rule shows also that $N_{S,t}$ fails to be closed under either conjunction introduction, or logical entailment, or, perhaps, both.

---

[36] Widerker 1987.

[37] O'Connor 2000, pp. 7-8. Two other attempted counterexamples by Anthony Kenny (1975, pp. 155-7) and Michael Slote (1982) are discussed and discarded by Fischer in Chapter 2 of his 1995 (pp. 25-45).

On the other hand, the failure of the closure principle is very counterintuitive and Widerker's counterexample is very peculiar. Maybe the counterintuitive result is the consequence of the weakness of the modal operator $N_{S,t}$, or more precisely the ambiguity about the exact content of it, that allows for weak readings, of which Widerker had taken advantage. Perhaps it is possible to sharpen the focus of the meaning of the operator, making the requirement expressed by it more exact, close to the strong end of the spectrum of readings that was allowed by the looser formulation, so that it fends off cases like Widerker's, yet still covers cases like that involved in the consequence argument.[38]

In a way it is like the case of the Gettier-type counterexamples to the justified true belief analysis of knowledge in epistemology. The analysis is very intuitive, Gettier's examples are quite *recherché*. It is not very clear how high, in terms of the degree of warrantedness against mistake, we should set the threshold for justification. The Gettier cases are dependent on a gap between justification and truth (as it will be discussed briefly in chapter 5). If we set the standards for justification *really* high, as for example in the case of Cartesian foundationalism, then there is no gap between justification and truth, and there is no room for Gettier cases.

This is exactly what O'Connor does in the first chapter of his *Persons and Causes* (2000).

$N_{S,t}$ (*p*) expresses the inability of person *S* at time *t* to make it the case that the state of affairs expressed by *p* does not obtain. In Widerker's counterexample it was interpreted as *S*'s inability to secure that not-*p* obtains. Sam was unable to secure that the bit of radium does emit a subatomic particle at t_2, although he was able to make it possible by refraining from destroying it at t_1. Similarly, Sam was unable to secure the falsity of the conditional whose antecedent was that the bit of radium does not emit a subatomic particle at t_2, and whose consequent was that he destroys it at t_1, because making a conditional false requires making its antecedent true while its consequent is false, but if the consequent is false, then Sam does not destroy the radium, and then he has no means to secure that it does not emit the particle.

But if we take the inability expressed by $N_{S,t}$ (*p*) to be stronger, so that it requires *S*'s is inability to perform an act (at t or later) that

[38] For a discussion of how the logic of different inability operators may differ, depending on their strength, see Kapitan 1996b.

would make *non*-p *as much as possible*, then this counterexample with Sam and the decaying radium is fended off. It is clear that, so interpreting $N_{S,t}$, neither $N_{Sam,t1}(p)$ nor $N_{Sam,t1}(p \supset q)$ are true, if p is the proposition that the bit of radium does not emit the particle at t_2, and q is the proposition that Sam destroys the radium at t_1. Sam can make it possible that the radium decays at t_2 (that is non-p) by refraining from destroying it at t_1, although he cannot ensure it. Similarly, he can make it possible that the conditional $p \supset q$ does not hold, likewise, by refraining form destroying the radium, which is making q false and non-p possible by the same token. So the obvious falsity of $N_{Sam,t1}(q)$ does not speak to the truth or falsity of the inference principle that $N_{S,t}(p)$ and $N_{S,t}(p \supset q)$ entails $N_{S,t}(q)$, at all.

So the dialectical situation is that we had a principle which had a very strong intuitive appeal, then someone came up with a clever counterexample, and then we pointed out the weakness in the formulation of the principle that made it vulnerable to that counterexample, fixed that point, and showed that the proposed example is not a counterexample to our principle in its fixed version. This is the dialectical point at which O'Connor leaves the discussion.[39]

Yet, I think, O'Connor's case can be pushed further. By strengthening the modal operator as he suggested we can fend off not just one or two particular types of counterexamples, but any counterexample purported to show that the inference principle from $N_{S,t}(p)$ and $N_{S,t}(p \supset q)$ to $N_{S,t}(q)$ doesn't universally hold. This is so because it can be shown that $N_{S,t}$, taken in the strong sense suggested by O'Connor, is closed under both conjunction introduction and logical entailment.

Closure under conjunction introduction would mean that $N_{S,t}(p)$, $N_{S,t}(q)$, and $\sim N_{S,t}(p \& q)$ form an inconsistent triad. They do.

$N_{S,t}(p)$ means that S cannot make it the case at t that $\Diamond \sim p$. $N_{S,t}(q)$ means that S cannot make it the case at t that $\Diamond \sim q$. $\sim N_{S,t}(p \& q)$ entails that S *can* make it the case at t that $\Diamond \sim (p \& q)$. But that entails that S can make it the case either that $\Diamond \sim p$, or that $\Diamond \sim q$. But that is a contradiction, since both of these are excluded given $N_{S,t}(p)$ and

---

[39] It is not quite true. In a footnote he also shows that the fixed principle is also immune to another kind of counterexample, which was offered by Kadri Vihvelin (1988) using a stronger inability operator than the one used in Widerker's example, yet not as strong as the one that figures in the fixed principle. (O'Connor 2000, footnote 14, p. 14.)

$N_{S,t}(q)$. So the triad is indeed inconsistent, therefore, $N_{S,t}$ is closed under conjunction introduction.

Closure under logical entailment would mean that $N_{S,t}(p)$, $N_{S,t}(p \supset q)$, and $\sim N_{S,t}(q)$ form an inconsistent triad. They do.

$N_{S,t}(p \supset q)$ means that S cannot make it the case at t that $\Diamond\sim(p \supset q)$. $\sim(p \supset q)$ is equivalent to $p \& \sim q$. So S cannot make it the case at t that $\Diamond(p \& \sim q)$. Supposing $\sim N_{S,t}(q)$, there is at least one thing that S can do at t that would make $\sim q$ possible. Suppose S does that thing at t. Doesn't he make also $(p \& \sim q)$ possible by the same token? There is only one way that he could fail to make also $(p \& \sim q)$ possible by doing what he does at t, and this is if this action of his would make $\sim p$ possible. Unless $\sim p$ is possible, whatever act makes $\sim q$ possible, surely makes $(p \& \sim q)$ possible, too. But whatever it is that S does at t, it surely doesn't make $\sim p$ possible, since that would violate $N_{S,t}(p)$, which says that there is no thing S could do at t that would make $\sim p$ possible. So this is a contradiction, the triad is inconsistent. $N_{S,t}$ is closed under logical entailment.

But we have seen earlier (the proof was given in footnote 35) that closure under conjunction introduction and logical entailment guarantees the validity of the inference from $N_{S,t}(p)$ and $N_{S,t}(p \supset q)$ to $N_{S,t}(q)$ universally. And that is what was needed to secure.

Coming back to the consequence argument, the validity of the inference from premises 1, 2, and 3 to 4 is what was drawn into question. Now it is an instant of the inference principle from $N_{S,t}(p)$ and $N_{S,t}(p \supset q)$ to $N_{S,t}(q)$, with p being the conjunction of a relevant description of the past and that of the laws of nature, and q being a relevant description of the present and the future including any arbitrarily chosen action of any arbitrarily chosen agent. Evidently, $N_{S,t}(p \supset q)$ is then premise 3 of the consequence argument, i.e., the thesis of determinism. The entailment from 1, 2 and 3 to 4 of the argument fits the scheme of the inference principle, the only remaining question is whether the inability we are confronted with in respect of changing the past or the laws of nature is the strong inability invoked by O'Connor, which made the principle waterproof.[40]

Surely we cannot make it the case now that it would be even as much as possible that a proposition about the past that has so far been true would be false from now on. The same applies to

[40] If it is, then N(1) and N(2) entails N(1&2), so it entails N(p), p being 1&2, so the entailment from 1, 2, and 3 to 4 really fits the scheme.

propositions expressing natural laws. So we can conclude that the inabilities that figure in the consequence argument are such strong inabilities.

I conclude that the attempt to undermine the entailment from 1, 2 and 3 to 4 in the consequent argument fails, thanks to O'Connor's scrutiny on the meaning of the modal operator involved in the premises. (It is a possible position to deny that O'Connor's strong inability is the inability relevant for freedom, but it is equivalent to denying the entailment from 4 to 5, to be discussed a few pages below.)

Attacking premise 2: the "fixity of the laws of nature"

Premise 2 has been challenged by David Lewis.[41]

Lewis does not deny premise 2 head on. He says, rather, that premise 2 is ambiguous between two possible readings. One is that we are incapable of breaking a natural law, that is, acting so that our action would itself be a law-breaking event or would cause one. The other is that we are incapable of acting so that, were we so to act, a natural law would be broken. Lewis concedes that the premise holds on the first reading of it, but claims that it falls on the second.

Lewis says it is not absurd to imagine that a "local miracle", perhaps immediately prior to our action, makes it possible that we act in a way that, without the miracle, would be a law-breaking event. Claiming that it is possible is nowhere near to claiming that we have the marvellous power to break the laws of nature. So premise 2 of the consequence argument, as it stands, is too strong. The correct version of premise 2 would state only that no action of ours can either be itself a law-breaking event, or cause one. That much is certainly true. But this weaker version of the premise, together with premise 1 (the fixity of the past), and premise 3 (determinism) does not entail that no one is free to do otherwise than one actually does.

Are miracles compatible with determinism? Maybe the occurrence of the miracle realizes a possible world which is different from what the actual world would be if the miracle did not take place. The miracle is a miracle only from this latter-worldly perspective. From the perspective of the world which is realized if the miracle obtains, there is no miracle. The world evolves as one would expect it on the

[41] Lewis 1981.

ground of the laws and the initial conditions. The laws are, or at least one law is, of course, different.

On the face of it, it seems like giving up determinism, since this picture involves a branching of the possible course of events. Isn't exactly this kind of branching which is ruled out by determinism?

There is not necessarily a branching. There is an alternative description of the situation according to which we have two distinct possible worlds all the way through. Although there is a perfect match in the course of events in the two worlds up to the moment when the miracle happens, they are two distinct and different worlds even before the miracle. They are not just numerically different. They are different also qualitatively, since their laws are different. That is why there is neither a branching nor a breaking of a law.

There are two grave problems with this picture, however.

The first is related to the "nature" of natural laws, so to speak. We think there are natural laws because we observe regularities. We come up with hypotheses that cover the regularities observed thus far, and some of these hypotheses stand the test of further observations. These surviving hypotheses are called natural laws. Laws are simply not to cover irregularities.

But Lewis insists that the miracle is not an irregularity. It is not really a miracle. It is a regular event, regular with respect to a different regularity, captured by different law. Different from what so far we might have thought it was.

But if we are free, in the sense that Lewis aims to secure by invoking the miracles, i.e. in the sense of having genuine alternatives, then the womb of the future hides myriads of events that are irregular relative to the laws we now think characterize our world. Are there alternative laws to cover all of them, plus everything which the laws we presently hold true have so far explained? Certainly there are no laws to cover just any arbitrary pattern of events. What guarantees that there are even two different coherent sets of deterministic laws that cover the series of events in two possible worlds which have been exactly alike for some 16 billion years and start to diverge only when I decide whether I should keep on typing now, or should rather go out and grab something to eat?

If there are laws to cover any arbitrary series of events then there are no rules really. Then no amount of empirical data would ever corroborate a causal law, to any degree, or would ever be a fair ground even for hypothesizing one. Any apparent regularity in

experience is then merely accidental. If laws can cover anything, they are worth of nothing.

The only alternative is to withdraw the claim that every time when a "miracle" happens it is only that a law is proven to be different from what we thus far though it would be. But here we face the a problem. If we don't hold on to the fantastic claim that, for all "miracles", there are laws that cover everything that happened before the miracle *plus the miracle*, then the "miracle" is a real miracle, and there is no room for inverted commas. But then every single miracle falsifies determinism.

But suppose we grant Lewis that laws can cover whatever needs to be covered, and then the inverted commas prevail and determinism is saved. Is the Lewisian compatibilist willing to pay the price that saving determinism this way costs? For the price is freedom, understood as freedom from the bonds of laws and initial conditions, the kind of freedom he wanted to rescue with his "miracles".

If the "miracles" are "miracles", with the inverted commas on, and not bare, determinism-falsifying miracles, then in performing the act that the "miracle" makes possible, the agent is as much the prisoner of laws and initial conditions as he would be without the miracle. It is just that the laws are different from what we have originally thought they were. The initial conditions are the same. He cannot do any other than what is prescribed for him by the familiar initial conditions and the less familiar laws.

I can think of only one last refuge for the local miracle compatibilist. Suppose that a crucial law is genuinely ambiguous before the miracle happens, and the obtaining or not obtaining of the miracle disambiguates it. It is not that the obtaining or not obtaining of the miracle decides whether the actual world is Possible World One, in which Reading One of the ambiguous law is the law, or Possible World Two, in which Reading Two rules. The ambiguity about the law has never been epistemic. It is really ambiguous, *in itself*, its ambiguity is a metaphysical reality before the miracle happens or not happens. Suppose further that the obtaining or not obtaining of the miracle depends on the choice of an agent. It is the choice of the agent that disambiguates the law. Had it been nonambiguous before the choice, there would have been no choice really. What the agent would choose would have been prescribed by the unambiguous law and the initial condition. But it doesn't get disambiguated prior to the choice. The agent is as free to choose as a libertarian free agent is.

Nevertheless, determinism is unbroken at least in one technical sense, because it is always true that only one course of events is logically compatible with the laws of nature as they are disambiguated by that course of event, and with the past.

This is a very exotic world, metaphysically speaking, in which sometimes laws determine events, sometimes the other way around. Laws change through time. They evolve from ambiguity to nonambiguity, as free agents make their choices. In every instant, only one state of the world is compatible with the past and the laws of nature *that hold then*. It is not that laws go out of fashion. No choice can ever falsify a law that was in effect that far. Choices only make them sharper. Nevertheless, it is not true that what has happened *so far* and the laws that hold now *before anything else happen* would entail what is going to happen. In this sense, and this is the relevant sense, determinism doesn't hold.

Challenging premise 1: the "fixity of the past"

The idea of "Multiple Pasts Compatibilism" is analogous to that of Lewis's "Local Miracle Compatibilism". It is true, the proponents of the idea say, that I cannot change the past, but there are actions which are in my power to do and which require that the past be different. It is not that I would have the power of initiating a backward-flowing causal chain. The situation is rather like this: Given the laws of nature, a backtracking conditional describes the relationship of a present action of mine and a past event: were I to do this-and-this, different from what I seem to be determined to do by the past and the laws, then that-and-that, different from what have actually happened, would have taken place prior to that. The multiple-pasts compatibilist asserts that the backtracking conditional and a "can-claim", i.e., the claim that it is in my power to perform the action in the antecedent of the conditional, can simultaneously be true. So premise 1 of the consequence argument, i.e. the principle of the "fixity of the past" is false on the "non-causal" reading of it, and that is the reading that is relevant.

John Martin Fischer discusses an alleged example for the simultaneous truth of such a pair of a backtracking conditional and a can-claim devised by John Turk Saunders.

> Suppose that I know that my friend believes that I will do X, and I am the sort of person who, in a situation like this, would not want to let down, and would not let down, a friend who believes that I am going to do X. Suppose that I am the sort of person who, in a situation like this, would want to refrain from X, and would refrain from X, only if my friend had not believed that I was going to do X. Then we may properly say that I would refrain from X, only if the past had been different, i.e., only if my friend had not held a belief that in fact he did hold. I have the power to refrain from X, and this is a power that I would want to exercise, and would exercise, only if the past had been different in that a belief that was held had not been held. So my power to refrain from X is a power so to act that (to perform an act such that if it were performed) the past would have been different in that a belief that was held would not have been held. And what is contradictory in this?[42]

Now I think I can tell "what is contradictory in this". There is a fairly obvious trade-off between the truth-values of the can-claim and the backtracking conditional, so to speak.

My ability to refrain from X in the example, if I have such an ability, is dependent on the relevant backtracking conditional being not exactly true. The truth of a somewhat weaker proposition than the backtracker in question is indeed compatible with my ability to refrain from X. It is almost as if my refraining from X would require that the past would have been different—which is equivalent to saying that it is almost the case that my refraining from X is impossible, as a change in the past *is* impossible. But it is not exactly impossible in the example that I refrain from X, only highly unlikely and very difficult. The can-claim is true only if it is psychologically possible for me to refrain from X, despite my loyalty to my friend. But if it is psychologically possible for me, then the backtracking conditional is false. It is not the case that were I to refrain from X, a different past belief of my friend would have taken the place of what he actually believed.

---

[42] J. T. Saunders 1968., quoted by Fischer 1995, p. 80. Fischer is also offering an analogous example of his own there taken form an earlier work titled "Power over the Past", Fischer 1984.

Or, alternatively, it may be the case that the backtracker is actually true. But then the can-claim is false. It might appear to me, on deliberating about whether I should X or not, that I can do both. But only one of these subjective options is genuinely available for me. From the other one I am locked away by psychological prohibitions which will guide my deliberation in such a way that there is no question about its outcome. I will do X. Of course, if we re-interpret the can-claim, so that it does not have to refer to an objective possibility, if it is O.K. if it only refers to a subjective alternative which I consider in the course of my deliberation, then, yes, the can-claim and the backtracker are compatible. But it doesn't speak to the question whether there are genuine alternatives if the past and some laws imply the future.

The simultaneous truth of such a pair of a can-claim and a backtracker would require the past be genuinely ambiguous until it gets disambiguated by the agent's making true the antecedent of the backtracking conditional, or by his refraining from it. In the above example it would mean that the past belief of my friend is genuinely ambiguous until I do X or refrain from it, and if I refrain from it, I make it the case that he did not believe that I would do X.

I try to imagine what the metaphysics of a world in which it is possible would look like.

Supposing that free actions are performed by a multitude of agents on a regular basis, in such a world there should always be an ambiguity about what happened so far, because otherwise there would be no room for further free actions. What it comes to, apparently, is that the truth-makers of at least some propositions about the past must be dispersed in time, so that some components of the truth-makers are in the future.

But how are they, then, propositions *about the past*?

I can think of only one metaphysical state of affairs which caters for something like this. It is that if some events are extended in spacetime.

They shouldn't be complex events, though. If they are construed from the events that happen at the particular points of the region of spacetime they occupy, then the constituent elementary events are the truth-makers of elementary propositions that have to be conjuncted in order to get the proposition that describes the complex event. The conjunction is made true when the last of the conjuncts is made true. There may be an ambiguity about the complex event up to that point,

and the complex event is partly in the past, so, in one sense, there is some ambiguity about the past. But there is no ambiguity about any of the past temporal parts of the complex event. For example, the event of my having a productive day today is such a complex event. The morning was quite productive, but maybe something distracts me from writing in the afternoon. There is an ambiguity about the complex event until then. There is an ambiguity about whether my morning today was the part of a productive day. But, being noon now, there is no ambiguity at all about anything that took place this morning, or about anything that depends causally on it.

So these genuinely ambiguous extended events should rather be *simple*. The idea of an extended simple event is that the spacetime-points that belong to it don't get filled in, so to speak, with anything, that is, nothing really happens in those spacetime points, until the whole extended event happens. The simultaneous truth of a backtracking conditional and the corresponding can-claim, would require that the antecedent of the conditional, the act what I'm now capable of performing according to the can-claim, would be the head of such an extended simple event in the present, and the consequent of the conditional would be part of the tail of the same extended event in the past. (It would also require that, against all appearances, the propositions that are the antecedent and the consequent of the backtracking conditional don't really individuate events.)

Happening is becoming, taking a place in existence. It is an ontological change underwent by the event. Once it went through this ontological change, it is fixed, so it is not a fit subject for changing any more. That is why "becoming" should be suspended until the whole extended event can come into existence, that is, until its temporally final point.

Looking from the outside, a world in which there are extended single events is not laid out in spacetime as a four-dimensional continuum of points. Rather, it is a mixture of four-dimensional points and four-dimensional spaghetti. It is in fact quite rich in spaghetti, because a piece of spaghetti corresponds to every free choice. The points, i.e., non-extended simple events can be thought of as the sauce on the spaghetti, if you like. Every piece of spaghetti is very long. Think of the piece of spaghetti, for example, that leads up to my doing now either A or B. The spaghetti should encompass everything that so far had to be one way or another backtrackingly depending on which way I choose to act now. If the causal fabric of

the world is as determinism suggests, allowing no forking in the temporal evolution in the world, then every piece of spaghetti has to go back to the Big Bang.

This world caters for the simultaneous truth of the backtracking conditional and the can-claim. But does it cater for a deterministic worldview?

In one technical sense determinism may hold. Spacetime points, or rather the contents with which spacetime points are filled in, are connected with deterministic laws both in the sauce and in the spaghetti pieces.

In the sauce, the normal way. But how can it be true within the spaghetti? Does it make sense to speak about deterministic laws that connect different parts of one single event that is extended in spacetime? Yes, we have to accept that this is possible. Remember, the backtracking conditionals are expressions of this connectedness within the spaghetti pieces. The truth of backtracking conditionals was part of the initial hypothesis. In this case, rather than laws that connect different events, we should speak of laws that connect fillings of spacetime points.

But at this point there are two possibilities we have to consider. Either the spaghetti is nomologically separated from the sauce, or it is not. If it is not, then there is a huge problem. Because then there is no spaghetti plus sauce really, there is only spaghetti. Because, were there also sauce, and were they nomologically connected, deterministic laws would secure a smooth transition on the border between the sauce and the spaghetti. Think of the infinitesimal calculus that made Newtonian mechanics possible. Reality must be smooth to be describable with differential equations. For the sauce to remain different from the spaghetti, no parts of the sauce should be initial conditions for time evolutions (solutions to the equations that express the laws) that have parts in the spaghetti. Because if the latter are ambiguous, so must be the former, if the laws are deterministic and the transition on the border smooth. So either the sauce is nomologically separate from the spaghetti, or it becomes spaghetti itself.

But the past cannot consist in spaghetti only, without sauce. Because then the past is empty, a clean sheet, as we normally think the future is. Then nothing really has taken place yet, because the chunks of spacetime occupied by spaghetti are filled in only when the

spaghetti ends. And spaghetties end in free choices that have not yet been made.

The only alternative is if the sauce and the spaghetti are nomologically separate. But in that world we could study only the laws of the sauce, and would never learn anything about the laws of the spaghetti. Any empirical data to verify a scientific hypothesis would be from the sauce, for the simple reason that in the spaghetti there is only empty spacetime and no data. We could hypothesise that the laws of the spaghetti are just like the laws of the sauce, but that hypothesis would be completely unfounded empirically. Logical positivists would say that such a hypothesis would be meaningless.

The only epistemic access to the laws in effect within the spaghetti would be available through spaghetti pieces that have already reached their endpoints, and whose endpoints are nonambiguous. It would require that their endpoints should not be connected with a backtracking conditional to any spaghetti that has not reached its end yet. But are there human choices that are not connected to any later human choice? Perhaps the last choice of a completely forgotten hermit who throws himself down from his rock is such a choice. So the study of dying hermits is the foremost source of knowledge about the laws governing our lives. Great. Can we study the last choice of the dying hermit without connecting it to open-ended spaghetties again?

Can anything be more absurd than this?

So this crazy world of spaghetti plus sauce is a world in which determinism (in one technical sense) and free choice are compatible. Backtracking conditionals and can-claims can be true simultaneously. But this is also a world in which it is very unlikely, maybe impossible, that we would ever learn anything about backtracking conditionals, i.e. the laws of the spaghetti. From the laws that would inform us about whatever is nomologically connected to human behaviour we are blocked epistemically. We wouldn't even know that there are true backtracking conditionals, because we would have no means to verify the hypothesis that spaghetti-points are deterministically connected. So it is a world which can be both deterministic and free, but in this world the consortium of scientist will never announce that the world is deterministic.

For the same reason, this world cannot be deterministic in exactly the sense involved in premise 3 of the Consequence Argument. As long as there is at least one human choice in the womb of the future,

there is a piece of spaghetti connecting that choice with the Big Bang. So it is not true that the past and the laws unambiguously entail the future, because the past is ambiguous. So we sunk premise 1 at the cost of adopting a ridiculous metaphysics, plus giving up hope that determinism will ever be made as much as remotely plausible empirically, plus giving up premise 3 as it stands. I am sure no determinist is ready to bear this cost.

I believe the absurdity of this suggestion is quite obvious as long as we operate on the assumption that time is an A series in McTaggart's sense.

But we have to take into account the possibility that time is not an A series, as the argument from the special theory of relativity suggests, and so there is no specific ontological status associated with pastness.

In that case we can have power over that past even in the stronger, causal sense. As the past has no ontological priority to the present, what determinism (in the above defined sense) comes to is only that one cannot choose a different action without choosing a world that is different throughout the whole of its temporal extension, but it is not the case that a large part of the world which is nomologically bound to a unique course of action is "already there".

Supposing the four-dimensional manifold to be ontologically homogeneous, no event is contingent relative to any previous event, there is no such thing as "contingency relative to what has already been laid down", yet, a whole chain of events stretching through the whole temporal extension of the world may be contingent in the nonrelational sense, meaning that the world, the four-dimensional block of events, would be different, were the agent to choose differently. There is a possible world, in which he chooses differently, and the source of this world's being the actual world, and not that one, is that he actually does not choose differently, although he could. This is not the power to *change* the past, because it is not the case that there is already something that needs to be adjusted to the choice. It is rather the power to *create* the past, together with the present and the future, one way or the other.

This is all familiar, but it has nothing to do with the examples of Saunders and Fischer. This is the Kantian picture which he advocates both in the *Critique of Pure Reason* and in the *Critique of Practical Reason* to reconcile libertarian freedom exercised timelessly in the noumenal world and determinism which holds true in the flowing time of the

phenomenal world in which the past and the laws of nature is consistent with only one way our choices can turn out:

> Now in order to remove the apparent contradiction between the mechanism of nature and freedom in the case under discussion, we must remember what was said in the *Critique of Pure Reason* or what it implies, viz., that natural necessity, which cannot coexist with the freedom of the subject, attaches merely to the determinations of a thing which stands under the conditions of time, and consequently applies only to the acting subject as appearance. As a consequence, [it pertains to the subject] only so far as the determining grounds of any action of the subject lie in what belongs to the past is no longer in his power; in this must be counted also his already performed acts and his character as a phenomenon as this is determined for him in his own eyes by those acts. But the same subject, which, on the other hand, is conscious also of his own existence as a thing-in-itself, also views his existence so far as it does not stand under temporal conditions, and to himself as determinable only by laws which he gives to himself through reason. In this existence nothing is antecedent to the determination of his will; every action and, in general, every changing determination of his existence according to his inner sense, even the entire history of his existence as a sensuous being, is seen in the consciousness of his intelligible existence as only a consequence, not as a determining ground of his causality as a noumenon.[43]

So the lack of future contingencies (relative contingencies) does not formally exclude libertarian freedom.

But it is difficult (although perhaps not impossible) to square the idea of libertarian freedom exercised tenselessly with the tensed phenomenology of decision-making. A theory would need to be supplemented of how it is that phenomenally my consciousness seems to crawl up along a worldline, how it is that I seem to know a lot of what is past and why does it seem correct to treat it as given,

---

[43] Kant 1788/1949, p. 203.

and, most importantly, why do I seem to make my decisions in the present and not timelessly, partly on the ground of the information I have about the past, and always in view of affecting the future and never the past. There must be a part of the past that is fixed because I have experienced it. There should be no way of changing it by my decision now, because that would undermine the trustworthiness of memory and the rational ground on which I seem to make my decision.

I don't think that a libertarian should be very keen on pursuing such a theory instead of pursuing ways of defending objective becoming and arguing against determinism. Yet, in a world without becoming it may, in principle, be true that large parts of the past, parts on which I do not rely as the given of my deliberation about my choice, are formed by my choice.

The picture may be even more complicated if we take into account that we are not the only free agents whose choices must be consistent with the past, and that parts of the past which do not figure in the given of my deliberation may figure in that of others. The free choices of all agents who make their choices partly on the ground of what has already taken place in their phenomenal time has to give out a single consistent history, and if free agents are numerous, if their knowledge of their past is rich, and if they interact frequently, then this constraint may effectively rule out any choice affecting the past even if the objective ontology of the world would otherwise allow for that.

It doesn't seem to be a promising line to take for a libertarian, yet, it should be noted that in a world without becoming premise 1 of the consequence argument can be questioned.

Is the inability operator closed also on the non-causal reading?

One of the common features of multiple pasts compatibilism and local miracle compatibilism was that both distinguished between two possible readings of premises 1 and 2, respectively, of the consequence argument. The two readings differed in the nature of the inability involved in them. They agreed that we are unable to bring about, either directly or causally, a different past or a different set of laws, but they insisted that we can do things that would require that either something in the past, or a law, would have been different. It is essentially the claim of the parallel assertability of a counterfactual conditional and a can-claim that asserts that the counterfactual

antecedent of the conditional we can make factual. The two readings of the premises, and of the inability they state, are often dubbed as the causal and the non-causal readings.[44] We have seen that both premises are true on both readings, as long as time is thought to be an A series. Earlier, building on O'Connor's work, we have seen that, on the causal reading, the inability stated by the premises is the kind of strong inability that secures the closure of the inability operator under conjunction introduction and logical entailment, which, in turn, guarantees the validity of the move from 1, 2 and 3 to 4 in the consequence argument. Now we have to check whether the same is true of the inabilities involved in premises 1 and 2 on the non-causal reading.

Is it true that no one can perform an action that would make it the case that the antecedent of a true counterfactual backtracking conditional would be as much as possible? Yes, because the metaphysical prerequisites for the backtracking conditional to be true and its antecedent to be possible are not different from the metaphysical prerequisites of the simultaneous truth of the backtracking conditional and the can-claim. They both require that the past be genuinely ambiguous until it gets disambiguated by the agent's making true of the antecedent of the backtracking conditional, or his refraining from it. As long as becoming is assumed to be a reality this is a metaphysical nonsense. So the inability involved in premise 1 is O'Connor's strong inability also on the non-causal reading.

Is it true that no one can perform an action that would make it the case that the antecedent of the true counterfactual conditional "*if the agent performed action A, at least one natural law would have been different*" would be as much as possible? Yes, it is true. We have examined the conditions under which the simultaneous truth of the counterfactual conditional and its antecedent would be as much as possible, and all conceivable ways of realizing those conditions led to consequences that were either absurd, or made the whole enterprise pointless from a compatibilist perspective, or both.

Challenging the entailment from 4 to 5

---

[44] See, for example, Kapitan 1996b.

Does it follow from that fact that we cannot make the present or the future different from what is prescribed by the past and the laws of nature that we could not do any other from the way we actually do? Is seems obvious that it does.

However, as it was mentioned in the previous chapter, it has been suggested that the ability to do otherwise than one actually does should be analysed with a counterfactual conditional: I could do otherwise = I would do otherwise, if I so willed.

Keith Lehrer argued, rightly to my mind, that conditionals fail to analyse the modality expressed by a "can" or a "could", because such modalities are probably irreducible. A sentence of the form "If C, then S X's" cannot mean the same as "S can X" for the former is compatible with "S cannot X", for example when Not-C is the case and Not-C entails that S cannot X.[45]

To this it can be objected that Lehrer may be right about one sense of "can" or "could" but this is not the sense that is relevant to the question of freedom. (Or, alternatively, one might claim that although O'Connor's strong inability operator is relevant if one sense of inability is concerned—the one related to objective modality—and in this sense the inability operator is indeed closed under conjunction and logical entailment, but this is not the sense that is relevant for freedom.)

This is a possible position. But if the defender of the conditional analysis takes this line, he admits that his theory is not that determinism is compatible with the objective existence of alternative courses of action, but that genuine alternatives do not really matter. What Lehrer says is that alternatives in the sense of the conditional analysis are not objective modalities. What the truth of the counterfactual conditional which has been suggested to analyse the can-do-otherwise-claim secures is not the objective existence of alternatives but the action's counterfactual dependence on the will, which is, in effect, the Hobbesian condition for freedom, i.e. that an action is free if it proceeds from the will.

It is uncontroversial that the determinatedness of the action by the will and the determinatedness of the will by prior causes is compatible with the existence of alternatives "in the practical perspective", i.e. with the introspective phenomenology of deliberation in which we do

[45] Lehrer 1968.

consider alternative courses of action as if they were genuinely possible.

However, as long as the question is whether the existence of alternatives *as objective possibilities* is compatible with determinism, despite the ingenuity of some compatibilist philosophers whose efforts were discussed in this chapter, the answer remains the obvious 'No'.

3 Is the Shallowness of Causal Control a Problem?

Two senses of self-determination

Being free partly means that our environment does not prevent our future from unfolding the way we would like it to unfold. A non-oppressive environment allows our future to be the expression of the internal forces in our personalities, such as the desires and the values, and the understanding of ourselves and the rest of the world, that we really embrace, that jointly manifest themselves in our actions. It is indeed a rich sense of freedom. It seems right to call it self-determination, since it involves the idea that our life is determined by ourselves rather than by anything else.

Self-determination we exercise by deliberating about how we should act. We canvas and consider alternative courses of actions and our life emerges (partly) out of the choices we make from these alternatives.

This conception of self-determination sits well with the causal conception of control. The special causes that determine our will, and with which we can identify or be identified by the relevant moral community, operate in the deliberative process. That these causes can be traced back to causes with which we do not identify, maybe to the distant past of the universe, does not matter according to the causal theorist. So this conception of self-determination is compatible with determinism. The alternative courses of action that we consider are not objective possibilities, they are possible only from our subjective perspective, whereas, in reality, laws of nature and initial conditions in the distant past of the universe may prescribe the whole deliberative process, including the range of alternatives that are to occur to us and the choice we make.

There is, however, another sense of self-determination.

None of us experiences what it is like to be everything one could possibly be. Being both a good scientist and a good philosopher is not within the reach of most of us. Committing ourselves fully to the pursuit of pleasures in the aesthetic sphere of life, and striving for perfection in the ethical sphere, in the Kierkegaardian sense, are incompatible projects. Dedicating our lives to the common good in public matters, and aiming for the most that is humanely possible in

taking care of our most beloved ones seem to be conflicting aspirations, too. Yet, if it is true that options not unlike these are *really* available to us, even if choosing one of them necessarily involves saying goodbye forever to the rest at some point or another, then it is a great thing about life. If the choice is really ours, especially if by choosing we decide about our one and only finite existence, then it is a great responsibility. This responsibility can be both a blessing and a curse, but I think most of us feel that our dignity as human persons would be significantly diminished if someone convinced us that it wasn't real.

If it is real, then freedom means more than just that our lives unfold from what we are and not from overriding outside influences. If it is real then we make a creative contribution to the coming about of something that is not fully contained in the present: our future self and our story.

This kind of self-determination is possible only if the libertarian conception of control makes sense and we can have genuine alternatives.

The introspective phenomenology of self-determination by deliberation and decision making does not decide the question whether we are self-determining in both of the above senses or only in the first sense. Introspectively we can see ourselves canvassing alternatives and weighing them against each other, but when it comes to the point when the decision is actually made the situation is much like Daniel Dennett described it in *Elbow Room*:

> We have to wait and see how we are going to decide something, and when we do decide, our decision bubbles up to consciousness from we not know where. We do not witness it being *made*; we witness its *arrival*.[46]

There are two groups of facts about this introspective phenomenology that may be relevant.

The first is that we did canvas and consider alternatives, did attend to and weigh reasons for or against choosing them, and the choice we finally made emerged from this process, whether or not determinism is true.

---

[46] 1984a, p. 78.

The second is that the alternatives have never been there objectively, if determinism is true, and when we are exercising (guidance) control, we are determined in this exercise by causes with respect to which we are not self-determining and with which we cannot reasonably identify or be identified.

The causal theorist must hold that only the first of these groups of facts is relevant for freedom. Daniel Dennett, for one, seems to argue in *Elbow Room* that given that we can have self-determination in the first sense, desiring also self-determination in the second sense is a philosophical mistake, a result of not thinking hard or clear enough. I take it that in his view this mistake consists of two components: a failure of appreciating how fully self-determining we are if we are self-determining in the first sense, and a failure to recognize that self-determination in the second sense is a confusion.

I postpone challenging the second component of this view until the eighth chapter. As far as the first component is concerned the significance of the two facts in the second of the above groups of facts is the question: the lack of objective alternatives and the shallowness of control on the causal account.

In respect of the lack of objective alternatives the core intuition of the causal theorist must be that we shouldn't worry about the nonexistence of alternatives that we don't want. The one that we want is a real possibility. There even seems to be an important relation between the fact that we want an alternative and the fact that it is a real possibility: that the latter is counterfactually dependent on the former. The causal theory of control, which is compatible with determinism, is compatible also with our never getting frustrated by the nonexistence of an alternative that we want.

Most of the discussion whether the lack of objective alternatives is a problem for freedom I leave to the next chapter about moral responsibility and rationality. However, in this chapter, which is mainly about the shallowness of control we will consider a scenario in which the shallowness of control is exploited in such a way that never getting frustrated by the nonexistence of an alternative that one wants is fully compatible with a strong and obvious kind of unfreedom.

For the discussion of what Dennett has to say about the shallowness of control on the causal account it is best to use Harry Frankfurt's hierarchical account of the will as a context, perhaps the subtlest of accounts that is compatible with the causal theory of control that has been produced to date.

The hierarchical account and absolutism about reflexivity

The hierarchical account is a theory of the mental faculty that controls action. This theory is not a species of the genus of causal theories of control, but it is compatible with the causal conception of control.[47]

The hierarchical account construes the freedom of the will as lying with the reflexivity immanent to our volitional capacities. This is essentially that we can reflect on whether we want to have particular desires or not. This reflexivity is the mark of our personhood, according to Frankfurt. This is what makes it possible that a person "is not only free to do what he wants to do; he is also free to want what he wants to want".[48] This is made possible, according to Frankfurt by a volitional structure that is complex and hierarchical.[49] Unlike a wanton, who is moved by its motivations without ever reflecting on them, we, as persons, may have second order desires about what first order desires we would like to have. We may act out the first order desires we approve on the second level, and deny those that we don't. Our will is free as long as we have this hierarchical volitional structure, and the particular will of ours that manifests itself in action is in conformity with our second and higher order volitions. So it is not only that we can act freely. We can have free will as well. A wanton can also act freely, but its will is not free. This account of freedom of the will makes no assumption on the causal ancestry of volitions. They may or may not be determined. So the freedom of the will, on this view, is compatible with determinism.[50]

---

[47] Frankfurt 1971. Frankfurt does not offer his account in the context of an explicit commitment to compatibilism. Nevertheless the account is presented as an alternative to the "prime mover unmoved" account of free will by libertarian philosopher Roderick Chisholm (cf. Chisholm 1966, p. 23). Frankfurt claims his account explains more and involves no "miracles" (p. 23). The hierarchical view of volitions advocated by Frankfurt has been extensively used to improve on classical compatibilism. For a bibliography on the significance of the hierarchical account for compatibilism see Kane 1996, p. 224, notes 1, 2, 4 and 5 to Chapter 5.

[48] Frankfurt 1971, p. 22.

[49] My attention has recently been drawn by Gábor Kendeffy to the fact that the idea of a hierarchical volitional structure and the freedom of the will conceived as stemming from the capacity of reflecting on whether lower order volitions are in line with higher order ones, was already present in Augustine's work, and can be found both in the first book of *On Free Choice of the Will (De libero arbitrio)*, and in the eighth book of the *Confessions*.

[50] When recapitulating the hierarchical account, I move back and forth between "volitions" and "desires" almost as if they were synonyms. They are, of course, not

The problem of shallowness in this content is that a person who has second and higher order volitions can be just as much a wanton with respect to these volitions as a plain wanton is with respect to its first order desires. Freedom in respect of volitions is construed with reference to how they are related to the subsequent level of volitions in the reflective hierarchy. But such a structure cannot accommodate an actually infinite series of ascents to higher levels. This problem was first pointed out by Gary Watson[51] and was later acknowledged by Frankfurt.[52]

Dennett, however, argues that it should not be considered as a vice.[53] He says the pushing for the freedom of increasingly higher

---

synonyms. But in the formation of a will to act, desires seem to be the critical things with respect to which we should want to be free. A volition can be formed either rationally or non-rationally. Suppose the difference is that in the latter case we are driven directly by a desire, while in the former beliefs about how the world is, e.g. beliefs about how certain desired ends are hooked up with certain possible means, or with certain desirable or undesirable consequences, may come into the picture. When the freedom of will-formation is the question, it is safe to speak about only desires, since we shouldn't long for freedom about beliefs in the same sense as about desires. *Prima facie* it wouldn't be a threat to our self-determination if our beliefs always automatically tracked the truth, and we couldn't help it. It doesn't mean, though, that when we reflect on the desirability of certain desires we wouldn't mobilize our knowledge about the world. When we speak of the hierarchy of desires, this contains an implicit reference to the epistemic input of will formation, which distinguishes between volitions and mere desires.

[51] Watson 1975.

[52] Frankfurt 1987, pp. 165-6.

[53] Dennett does not explicitly discuss the hierarchical model. Nevertheless in Chapter 2 of *Elbow Room* he uses the very idea of reflexivity to distinguish between sophisticated deterministic deliberators, which on his view are good candidates for being both conscious and rational, from an obviously non-conscious and non-rational deliberator, a wasp called *Sphex*. Dennett suggests that "the capacity for conscious recognition of motivations is…a necessary condition of real freedom", and also that reflection entails consciousness, because "the gulf between unconsciousness and consciousness has already been crossed once we have arrived at systems that are capable of treating some of their own internal 'belief' and 'desire' states" since no system can be "unconsciously self-conscious" (pp. 36-7). Notice that if we substitute "reflexive (hierarchical) volitional structure" in place of "capacity to conscious recognition of motivations" and "freedom of the will" for "real freedom" then Dennett's claim about freedom turns out to be almost word by word the same as Frankfurt's (1971). In the first section of his 1987 Frankfurt also suggests that reflexivity is the clue to consciousness, although the claim he makes is more modest than Dennett's. While Dennett hints that reflexivity is not just necessary but also sufficient for consciousness, Frankfurt treats reflexivity only as a necessary condition (pp. 160-2). This claim is akin to Frankfurt's earlier claim that "one essential difference between persons and other creatures is to be found in the structure of a person's will" (1971, p. 12), which is followed by the identification of reflexivity as a mark of personhood. What I am saying is that, although it is not explicitly said, Dennett does endorse at least the core of the hierarchical view. I am also confident that the views

order volitions is wrongheaded. If we ever want to act on our reflective deliberation, it had better be finite. Had we ever tried to take our absolutism about the freedom of the will seriously, we would have experienced what it is like to be paralyzed. If freedom has anything to do with the ability to act, the existence of some fixed, unreflected items in one's volitional structure is not an obstacle but a necessary condition for it to work. Reflection must stop somewhere. So the regress just stops, too, naturally when it hits these fixed items. Thank God, they are there. Where do they come from? They evolved.

Is the will free, if Dennett is right?

It is certainly not ultimately free in the sense that we might have had in mind when we decided to keep on inquiring about the freedom of volitions of increasingly higher orders. Nevertheless, it is at least freer than that of a wanton. Dennett's argument seems to substantiate that, although we may not be ultimately free on these terms, we may well be as free as it gets. Admittedly, the further we get with reflecting on our desires and commitments, the freer our will is. But given that the function of the will is to control action, and given that excessive reflection hinders control, not speaking of infinite reflection, which is equivalent to completely forfeiting action, it is not unreasonable to say that, after a certain point, any further gain in terms of the freedom of the will is always accompanied by a greater loss in terms of the freedom of action. The finiteness of our volitional hierarchy may be optimal for our "overall freedom", which is construed as an aggregate measure of the freedom of our will and the freedom we exercise by volitionally controlling our activity.

What Dennett leaves out of consideration, however, is that the alternative that the libertarian conception of control offers to the shallowness of control on the causal conception of it does not require an actually infinite series of reflections.

To be sure, in order to contribute to the actual real-time volitional control of activity, deliberation must be finite. Dennett rightly observes that it rules out the possibility of an actually infinite series of reflective ascents in the hierarchy of desires. But it is far from clear that whoever is an "absolutist" about the freedom of the will has to be committed to an actually infinite series of reflections.

I attribute to him in this paragraph do follow from what he says at various places in *Elbow Room* such as the first paragraph on p. 70, pp. 108-13, top of p. 119, pp. 164-5, and others.

We haven't yet specified what this "absolute" freedom would exactly consist in. So far it was only assumed that "absoluteness" in respect of the freedom of the will requires that all volitions at all levels be free.

Even if we accept that conformity, or the *lack of conflict* with higher-order volitions is a necessary condition for the freedom of volitions, it does not entail that it is also a necessary condition that the conformity be actually checked. Only if this latter was also necessary would absolutism about the freedom of the will require an actually infinite series of reflective evaluations.

One might rightly object, however, that reflexivity had a more fundamental role than the existence of a hierarchical structure in the intuitions that led to the Frankfurtian account of free will. The hierarchy of volitions came into picture only to make adequate provisions for reflexivity.[54] The mere existence of conformity with higher-order items in the hierarchy of volitions without reflexivity does not seem to yield freedom.

Keeping this in mind, wouldn't it be enough for the freedom of the will to require only the *capacity of reflexivity* and not also it's being actually exercised? What I mean by that is the capability of suspending identification with a desire any time it looks necessary, taking a step back, and drawing in question our commitment even to the desires of the highest level that has so far been reached by reflection, if we have any doubt about them. Of course, it must be accompanied by requiring also a kind of sensitivity to conflicts between desires of different orders if they happen to lurk in the background. Otherwise even a wanton that is, in principle, capable of reflection, but is always doubtless about its desires would easily satisfy the proposed condition for the freedom of the will. This kind of absolutism is not a paralyzing one. For absolutism about the freedom of a hierarchically structured will so understood requires only that adding one more round of reflection to the process of deliberation be always possible, regardless of how many rounds of reflection it has already involved. This, however, is only potential infinity, which is easy to reconcile with the practical wisdom that deliberations always need to be actually finite.

The escape from the regress problem that Frankfurt himself proposed in an essay which he offered in 1987 as an improvement on

---

[54] This is emphasized also by Frankfurt: 1987, p. 165, note 7.

the original 1971 statement of the hierarchical account, follows similar lines. He hopes to escape the regress by an appeal to the notions of *identification* and *wholeheartedness*.[55] The notion of identification was already present in the 1971 paper:

> When a person identifies himself *decisively* with one of his first-order desires, this commitment 'resounds' throughout the potentially endless array of higher orders. … The fact that his second-order volition to be moved by this desire is a decisive one means that there is no room for questions concerning the pertinence of volitions of higher orders. … The decisiveness of the commitment he has made means that he has decided that no further questions about his second-order volition, at any higher order, remain to be asked.[56]

Watson argued that this appeal to identification is uninstructive about how we avoid wantonness with respect to higher-order volitions. It seems that the "decisive commitment" followed by a "resonance effect" means simply that the interminable ascent to ever higher orders is just not permitted, and this is arbitrary unless some additional account is offered of how and why the relation between the person and the desires he is decisively committed to is so special.[57] In the 1987 essay Frankfurt explains the resonance effect in terms of the belief that no further inquiry could override the commitment:

> For a commitment is decisive if and only if it is made without reservation, and making a commitment without reservation means that the person who makes it does so in the belief that no further accurate inquiry would require him to change his mind. It is therefore pointless to pursue the inquiry any further.[58]

---

[55] Hence the title of the essay: *Identification and Wholeheartedness* (1987).
[56] Frankfurt 1971, pp. 21-2. I cite Frankfurt with the same italics as he cites himself in his 1987 (p. 167).
[57] Watson 1975, p. 218.
[58] Frankfurt 1987, pp. 168-9.

If there is no apparent conflict between desires of either the same or different orders, nor has the person any reason to suspect that such a conflict may be uncovered by further reflection, then terminating the reflective sequence is not arbitrary. Then

> the person no longer holds himself apart from the desire to which he has committed himself. … To this extent the person, in making a decision by which he identifies with a desire, *constitutes himself*.[59]

Ordering of competing desires and conflict resolution between mutually exclusive ones is the way "to create a self out of the raw materials of inner life" (p. 170). Making a decision is not "a simple act that merely implements a first-order desire", because "it necessarily involves reflexivity, including desires and volitions of a higher order" (p. 176). Persons who resolved the conflicts between their desires enjoy a state of volitional unity called *wholeheartedness*. They are not ambivalent about what they want and what they want to want. They have the will they want to have. So their will is free.

So says Frankfurt. But should the "absolutist" be satisfied? I think it depends on how it happens that "the person no longer holds himself apart from the desire to which he has committed himself". The absolutist should require that holding himself apart from the desire in question and further inquiring about conformity with higher order volitions be a living option for the person, even though he does not exercise it. Abstaining from further reflection is done "in the belief that no further accurate inquiry would require him to change his mind". The ascent to ever higher orders in the volitional structure should not be *actually infinite*, that much is fine with the absolutist. Nevertheless, he should require it really be *potentially infinite*, meaning that ascent to one level higher should always be possible. But if the person's belief that no further reflection would make him change his mind is necessitated by factors fully external to his personality (by the state of the early universe plus the laws of nature, for example), then further reflection has never been possible, and abstaining from the decisive commitment has never been an option. So the absolutist may consent the wholeheartedness account of acting and willing freely, but not without qualifications, and with those qualifications the

[59] Ibid, p. 170.

wholeheartedness account is no longer compatible with determinism and requires the libertarian conception of control.

To sum up: An "absolutist" is a theorist who is worried about the shallowness of control on the causal account and prefers to avoid it. Dennett practically says he should not be worried, because the opposite of shallowness would be an infinite ascent to ever higher orders of the Frankfurtian hierarchical structure of our volitional faculty, which would be paralyzing. Frankfurt himself offers an account of the freedom of the will in terms of a state when this ascent to higher orders is terminated in the belief that no further reflection would bring to light conflicts within this structure. This is fine with the absolutist as long as that belief was adopted "freely" in the sense that it was objectively possible for the agent to keep on reflecting, he was not determined to terminate his reflection, i.e. freely in the libertarian sense. This way shallowness is of course avoided. My point is that if the absolutist gets what he wants, it does not paralyze the agent in any way, since it does not require an actually infinite series of reflections, it only requires that the series of reflections be potentially infinite, that is continuable at any point. So Dennett is wrong.

Walden Two

Walden Two is a utopian community. Its members enjoy an extraordinary freedom. They can do and have whatever they want. How is that possible? The founder of the community, a man named Frazier, organized it that they be conditioned from their early childhood to want only what they can do or have.

Benjamin Skinner, whose imagination created Walden Two, thought it the way to maximize human freedom.[60] Robert Kane hinted that Walden Two is a good way to maximize compatibilist freedom.[61]

Kane argued in *The Significance of Free Will* that the people of Walden Two are free on Harry Frankfurt's hierarchical account of freedom.

Taking into account of what has been said in the last section I would qualify this claim to be true of the hierarchical account in combination with the view that control is a species of causal determination, in consequence of which shallowness of control is

[60] Skinner 1962.
[61] Kane 1996, p. 66.

unavoidable, which, in the context of the hierarchical account means that reflexivity, even as a potential, breaks down at some points of the volitional hierarchy.

Walden Two seems to be a scenario that exploits the shallowness of control on the causal conception of it in such a way that it shows in respect of all varieties of freedom based on the causal conception of control that our intuitions rebel against accepting them as proper analyses of freedom.

Causal theorists (compatibilists) standardly present freedom as the freedom from constraint and coercion. This tradition goes back at least to the 17th century, to Hobbes, who, in his famous debate with Bishop Bramhall, argued that this is the freedom we normally recognize and desire in everyday life, and that it is compatible with determinism.[62] A modern, but still classic, proponent of the same idea is Ayer. Whenever I go through a process of deliberation to be followed by an action, and no exogenous factors, like a pistol pointed at my head by another agent, a paralysis of my limbs, or some kind of compulsive neurosis, influence my action, then, so says Ayer, I act freely. The fact that the result of my deliberation is predetermined is not a constraint, or only in a metaphoric sense.

> For it is not when my action has any cause at all, but only when it has a special sort of cause, that it is reckoned not to be free.

and

> It is because of the metaphor, and not because of the fact, that we come to think that there is an antithesis between causality and freedom.[63]

It is plain that there is absolutely no constraint or coercion in Walden Two. That the members of the community are conditioned not to want to do what they cannot do is not a constraint, or only in the "metaphoric sense".

As we have seen, central to the causal conception of freedom is the idea that our action is free as long as we act as we will. This idea can also be traced back to Hobbes. In the *Leviathan* he says

---

[62] Hobbes 1962, p. 35.
[63] Ayer 1954, pp. 21-2.

> The actions which men voluntarily do…because they proceed from their will, proceed from liberty; and yet because every act of man's will, and every desire and inclination proceedeth from some cause, and that from another cause in a continual chain…proceed from necessity.[64]

This view has been a standard element of compatibilism ever since.

The requirement that a free action should proceed from the agent's will is met with great consistency in Walden Two. It is never the case that what Walden Two people do is not what proceeds from their will. Because in that case they would do something different from what they want to do. But that is impossible, because they want only what they can have, so they never get frustrated. So if this is the condition of freedom, no one is freer than them.

In response to the worry that if determinism is true, we could never have done any other than what we actually did (or to the worry that if the causal conception of control is true we could have done other than what we actually did only in virtue of some random event turning out some other way than it actually did), compatibilists (causal theorists) classically proposed the conditional analysis of the ability claim, could have done otherwise = would have done otherwise, if so willed[65], as it was mentioned in the first chapter and briefly discussed in the last section of the second, that can be traced back to Hume's conception of "hypothetical liberty":

> If we choose to remain at rest, we may; if we choose to move, we also may. Now this hypothetical liberty is universally allowed to belong to everyone who is not a prisoner and in chains.[66]

The conditional analysis is still influential.[67] It is not a trivial matter how to conceive of the truth-conditions of a counterfactual conditional, which I will not enter here. Yet, I think, it is intuitive that

---

[64] Hobbes 1958, pp. 71-2.
[65] Moore 1912.
[66] Hume 1975, p. 104.
[67] Cf. Bok 1998.

in Walden Two if someone had willed otherwise than he actually did, it would have resulted in his doing otherwise than he actually did, otherwise the main principle of Walden Two, that everybody is conditioned to will so that his will could be fulfilled, would be broken.

There is little doubt that the inhabitants of Walden Two also enjoy the prestigious status of personhood in Frankfurt's sense. They are people just like us, so if we have a hierarchical volitional structure, they have it too. Their uninterestedness in purposes they could not successfully pursue must be supported by the entire hierarchy of the desires they commit themselves to, otherwise they would sometimes want things they could not have. At least they would higher-order-want to have lower order desires that cannot be theirs. This is impossible in Walden Two *ex hypothesi*. The conditioning they received guarantees that Walden-Twoers are always "marvellously wholehearted" in Frankfurt's sense, as Robert Kane observes, have the wills they want to have, and so what Frazier claims of Walden Two, i.e. that it is "the freest place on Earth"[68], comes out true on Frankfurt's theory, as "they have maximal freedom of will and action in the hierarchical sense".[69]

The point is that people who are perfectly free on conceptions of freedom that involve the causal conception of control, since that conception of control is shallow in the sense specified in the first chapter, may at the same time be victims of what Kane calls "covert non-constraining control", in which the controllers do not achieve their goal by constraining or coercing others against their will, but rather by manipulating their will so that they willingly do what the controllers want them to do.

Even if the causal conception of control is true and we can be self-determining only in the first of the two senses of the word discussed in the beginning of the chapter, there is a huge difference between acting under constraint or coercion and acting out of one's will. Even if determinism is true and our actions can be given an exhaustive causal explanation, there is a huge difference between different sorts of causes. Even if the causal chains that lead up to our deeds link everything we do unambiguously to the lifeless past of the universe, those chains go through us and the result carries the stamp of who we are. If the structure of the part of the causal machinery

---

[68] Skinner 1962, p. 297.
[69] Kane 1996. p. 65.

which is internal to us reveals sufficient depth, complexity, and reflexivity, then there is more than a good hope that we can make sense of the distinction between acts and mere happenings, and we can hold ourselves free in a sense "worth wanting", and we can sensibly ascribe to ourselves opportunity, choice, autonomy, creativity, desert, responsibility, individuality, dignity, hope, love, reason and the like. I do not question this.[70]

People of Walden Two, however, are maximally self-determining in this sense of the word. This, I believe, is enough motivation to want also self-determination in the second of the senses discussed in the beginning of the chapter, the one that requires control in the libertarian sense and genuine alternatives.

Is Walden Two an unfair intuition pump?

I suppose Dennett would protest that Walden Two is an "unfair intuition pump" and it is being abused here. Our intuitions suggest that the members of the community are not free because we are aware that they have been manipulated by Frazier and not because there would be anything inherently wrong with the compatibilist account of freedom of action and freedom of the will that has been offered. We are trying to make a Frazier out of determinism or of the shallowness of the causal conception of control, thereby contributing to the zoo of "bugbears" and "bogeymen" depicted in the first chapter of *Elbow Room.*

Dennett must believe, if he is honest in that chapter, that incompatibilism about freedom is motivated by a *fear* of determinism, which is fed partly by simplistic analogies just like the present one between it and Frazier. I think, to the contrary, that incompatibilism, or better, libertarianism, is motivated by a *hope* that there is something more to liberty, over and above the liberty that we can have if determinism holds, which is also worth wanting and can be ours if determinism is false. This extra would be that at least sometimes the ultimate origination of our acts lies with us, involving that no sufficient causal conditions for these acts of ours are traceable in the rest of the world, or in the history of the world,[71] we are the sole and

---

[70] That compatibilist freedom is "worth wanting" is a Dennettian slogan. As long as only this much is claimed, I agree.

[71] Dennett and Christopher Taylor in an essay titled "Who's Afraid of Determinism" (2002) argue that wanting ultimate origination is a literal confusion. Those who think

ultimate authors of our goals and purposes, at least to the extent of an act of approval and commitment (in the sense of judging that although further reflection and revision would be possible, it is not necessary) whose ultimate origination lies with us, and thereby, to the extent these commitments are self-constituting in a way envisaged by Frankfurt, we are the ultimate and non-derived shapers of our selves.[72]

Maybe this hope is doomed to fail, either because its object involves some internal incoherence, or for some empirical reason. Determinism would be one such reason. Our desires and commitments being manipulated by a behavioural scientist would be another. The libertarian concern is whether our hope stands or falls, and if it falls, it is of little interest whether it falls for a reason which itself is well worth to fear (like a superscientist in control), or for a neutral one (like determinism).

If so far I failed to convince, let me use another intuition pump. Suppose that Walden was not a lake somewhere in the woods of 19th century New England, but an Island. Or, better, let this Island be called Walden One, so it is distinguished from Thoreau's original Walden. Suppose further that this island had its own way in biological evolution, which was different, although not very different, from the evolution on the mainland. The peculiarity of the evolution on Walden One is that on this island not wishing what one cannot have or do has an extremely great survival value. Surely, it has some on the mainland (in the real world), as well. What we need to imagine is only, that it has significantly more on Walden One. Now suppose that the first primates on the island originally had the capacity of "ultimate

---

they should want it confuse necessary causal conditions with sufficient ones. Our causal significance in bringing about our acts is dear to our hearts, and it is O.K. But it requires only that we be necessary causal conditions for the obtaining of our acts. The fact that there are sufficient causal conditions in the remote past of the universe for our acts to obtain is fully compatible with our contribution's being necessary. Libertarians think determinism is harmful either because they are unable to see the distinction between necessary conditions on the one hand and sufficient conditions on the other, or because they are sloppy enough to speak about only 'causes' never bothering with the distinction between what causal conditions are sufficient and what are necessary. Of course, this is not so. Of course what libertarians want is that there be no sufficient causal conditions for our acts in the remote history of the universe. I don't think accusing the other party with failure to recognize obvious and elementary distinctions and the like would promote the quest for the truth in the issue of freedom, or in any issue, for that matter.

[72] This is the same as to require that at least sometimes we would be able to perform self-forming actions in Kane's sense (1996), and this is the same as what Kane calls UR (Ultimate Responsibility) condition, or Martha Klein U-condition (Klein 1990).

origination" (in the sense specified above). They also had the capacity to respond to the environment with changes in their deliberative faculty that inhibit deliberative processes that would lead to volitions that could not be carried out successfully. This latter capacity did not always work reliably, because it was sometimes overridden by the former. Tens of thousands of years of evolution, however, made some of their descendants (who are otherwise exactly like us) much more effective deliberators. They were selected for the latter capacity (avoiding frustration) all the way through. The capacity of "ultimate origination" died out, or rather, domesticated: the inhabitants of the island form their wills freely, but only relative to a fixed pattern of some unquestionable desires and purposes, whose sole function is to secure that nobody ever wants to do or have anything he couldn't do or have under the given conditions. Continental fishermen who every once in while sail close to it call the island in their tongue The Island of Unspeakable Peace. Of course, nobody believes the stories they tell of the Blessed People that dwell there. Nobody, except one strange fellow from a nearby university, an eccentric and somewhat arrogant behavioural scientist called Frazier.

Now, suppose that Frazier is absolutely benevolent. He honestly believes that the Blessed People are really blessed, and he honestly regrets that he is not one of them. What he wants is only to make their freedom, happiness, and peace of mind attainable to mainland people. So he tries nothing more in the Walden Two experiment than achieving the same result with ordinary people by way of conditioning them from their early childhood.

Are, then, the Walden One islanders better off than the inhabitants of Walden Two? Is it really credible that the people of Walden Two are unfree because the limits to the freedom of their will came from the wrong source, while the people of Walden One, with exactly the same limits, are free (indeed, maximally free, as it is on the causal account), only because their limits had naturally evolved? Is maximal freedom really discriminated from unfreedom by a condition which is fully external to these people's ways of doing things, is it credible that the distinction between them has nothing to do with how their deliberating faculties are hooked up, but only with who or what hooked them up so?

I think if Walden One is a free place, so is Walden Two, and vice versa.[73] If our judgement is that Walden Two people are not free, this is not because Frazier is around. Frazier's presence *only makes it obvious* that their self-determination is compromised.[74]

[73] It may be argued that Walden One and Walden Two are not exactly on a par. In Walden One the capacity of ultimate origination died out or degraded irreversibly, whereas in Walden Two only the strong habit of not using it, or using it in a very constrained way has been developed, so Walden-Twoers can in principle be waken up, so to speak, but the "blessed people" cannot. I think the disanalogy only makes my case stronger: if there is a difference in the degree of freedom between the two communities, Walden-Twoers are better off, although the "bogeyman" is around.

[74] It seems that this is a rare occasion when John Martin Fischer and I are in agreement. See his discussion of Dennett's "bugbears" and "bogeyman" in Section 4 of the first chapter of *The Metaphysics of Free Will* (1995, pp.14-21).

4 Do We Need Genuine Alternatives? – A. A Letter from Conrad

Conrad is a character in Dennett's recent book about freedom. Conrad is relatively bright, but not exceptionally sharp. He thinks that the lack of genuine alternatives is incompatible with freedom and moral responsibility. He is stuck in common sense. Whenever Dennett answers anticipated objections to his views, the objector is Conrad throughout the book. I am not sure Conrad gets a fair treatment. I try to fight back on his behalf.

The reconstruction of Dennett's main line of argument to the effect that determinism and responsibility-conveying freedom are compatible

Dennett's main aim is to show, or at least make it plausible, that the yet to be achieved complete and exhaustive scientific view of ourselves, on the one hand, and our commonsense view of ourselves as morally responsible free agents, on the other, are compatible. Moreover, they will remain compatible even if the scientific view of ourselves, when completed, turns out to be deterministic. Or indeterministic.

So Dennett thinks that we shouldn't worry that either determinism or randomness will spoil freedom, because freedom and responsibility do not turn on the causal organization of the world. They turn on a clever design that makes it possible that some organisms are free and responsible. It is not only an individual but also a social design. The design principles that make freedom and responsibility possible are equally realizable in a deterministic or in an indeterministic world. And they don't presuppose an intelligent designer. They could have evolved.

Both books in which he advocates this thesis, *Elbow Room* of 1984 and *Freedom Evolves* of 2003, are written in a very informal fashion. Perhaps the best way to decide whether we find his reasoning compelling is to attempt a somewhat formalized reconstruction of his main line of reasoning.

The central claims made in Dennett's two books are these:

1 A material mechanism, deterministic or not, can be correctly said to be a rational agent, that is, it can be correctly said to act, have reason, and act on reasons.
2 A world whose future is already laid down *sub specie aeternitatis* (and a deterministic world would be such), can accommodate possibilities that are not actualities. A deterministic mechanism can be correctly said to have opportunities.
3 A material mechanism can be an author of decisions and not just the locus of the causal summation of external influences and/or random occurrences that are relevant for the occurrence of his actions.
4 A material mechanism, even if it is deterministic, can be correctly said to be responsible for its actions in a morally relevant sense.
5 (Therefore) A deterministic mechanism can enjoy a freedom which is worth wanting and which conveys responsibility.
6 An agent whose design involves indeterministic parts can also be free in the same sense. However, an undetermined agent cannot be freer, or free in a fuller sense, than a deterministic mechanism, independently of the truth or falsity of determinism.
7 (Therefore) Whether the causal fabric of the world is deterministic or involves genuine indeterminacy has no bearing whatever on the problem of freedom. We can be free either way.

Once this much has been established, Dennett's compatibility thesis almost certainly comes out true. Some think the complete scientific view of ourselves is incompatible with freedom because they think the scientific view involves determinism (at least on the macroscopic level) and freedom requires (macro) indeterminism. Some compatibilists about freedom and determinism think that freedom requires determinism, because indeterminacy in the causal genealogy of action would put us in the mercy of pure chance. Claim 7, if true, proves both parties wrong. Neither determinism nor indeterminism is a precondition for freedom and responsibility. So the main obstacles are cleared away from the way of the compatibility thesis. What is required for freedom and responsibility is rather a design that makes knowledge about the environment, anticipation, representation of possible courses of action, and a rational choice between them possible. Such a design can be realized by material mechanisms, whether they are deterministic or include indeterministic parts. Such a design can be a result of evolutionary selection. So there

is no prospect for completing science, in particular the scientific explanation of our actions, in a way that it would be incompatible with freedom or responsibility.

Dennett's argument is a defence of what we earlier called the causal conception of control and freedom. On that conception control was the determination of the will by a causal mechanism of some specific sort, of one specific sort of design, one can say.

Claim 5 is supposed to follow from Claims 1-4. Claim 7 is supposed to follow from Claims 5 and 6. Claims 1-4 and 6 are supported by independent lines of reasoning in the two books.[75]

To be sure, the sequence of Claims from 1 to 7 was not supposed to be a properly formalized reconstruction of the Dennettian argument. The logical relation between Claims 1-4 and Claim 5 is not an instance of strict entailment. That would require an additional premise about the exact conditions for freedom "worth wanting" to be predictable about an agent. In relation to this, Claims 1-3 would also need to be made more precise. More would need to be said about the exact sense in which agency, rationality and authorship is predicated of deterministic agents, and about the sense in which a deterministic world is claimed to accommodate possibilities.

But I am sure Dennett could easily provide these additions. That his argument is not formalized is not necessarily a vice. It is not this on which the success of the argument turns. I shall not debate that the entailment relations hold. Nor shall I debate Claims 1-3. In the more formalized version these premises would involve qualifications, which would refer to the distinction between the "practical" and the "theoretical perspectives"[76], or between the "intentional" and the

---

[75] Claim 1: Chapter 2 *Elbow Room*; Chapter 2 (action) and Chapters 5 and 9 (rationality) *Freedom Evolves*.
Claim 2: Chapters 3 and 5 *Elbow Room*; Chapter 3 *Freedom Evolves*.
Claim 3: Chapter 5 *Elbow Room*, Chapter 8 *Freedom Evolves*.
Claim 4: Chapter 7 *Elbow Room*, Chapters 9 and 10 *Freedom Evolves*.
Claim 6: Chapter 5 Section 3 *Elbow Room*; Chapter 4 *Freedom Evolves*
Claim 6 is also argued for in Dennett 1995. In a note to this essay Dennett says what he thinks the problems of freedom are. In the wording of the above Dennettian claims I followed closely the wording he gives to these problems, as if the claims were the answers to the problems, in order to be as faithful as possible to both the spirit and the word of his thought.

[76] Which is already familiar from previous chapters. See the discussion of the distinction between the two perspectives drawn by Bok in her *Freedom and Responsibility* on p. 84.

"physical stances".[77] I don't doubt that there is a sense in which agency, rationality and authorship can be predicated of deterministic material mechanisms, and that there is a corresponding sense of freedom that requires these things only in these senses. I concede that freedom in that sense is worth wanting.

I have grave doubts, however, about Claim 4, and I think Claim 4 is the element of the argument on which its success turns. Dennett offers many reasons for this claim both in *Elbow Room* and in *Freedom Evolves*, but I find these reasons highly unsatisfactory. Similarly I feel towards Claim 6. But Claim 6 falls automatically if it turns out that a deterministic agent cannot be morally responsible but an indeterministic agent can, because the sense of freedom that conveys responsibility is fuller than the one that doesn't. So Claim 6 is not independent of Claim 4. And the argument falls anyway if Claim 4 falls. So I shall concentrate on Claim 4.

But before starting the discussion of Claim 4 a word of clarification. Of course, I will *not* try to show that Dennett's main claim about the compatibility of freedom and responsibility *with science completed* is false. I do hope it is true. We don't know yet what science, when completed, whatever that means, will say about the causal fabric of the world. I hope it will say that the world is indeterministic. That is why I hope science completed will be compatible with freedom and responsibility. Surely, I don't want to refute this. The compatibility claim I am trying to refute is the one contained in Claim 4, the one that concerns determinism and moral responsibility. And, through that, I try to refute the claim that the truth or falsity of determinism is irrelevant for freedom. That is Claim 7, the conclusion of the above reconstructed argument.

The two readings of "correctly" in the critical premise

Claim 4 was that deterministic material mechanisms, under specifiable circumstances, can be correctly said to be morally responsible for what they do. I would like to highlight the adverb "correctly" in this sentence. I think "correctly" in this context can mean two things. In case responsibility is an objective matter of fact, a metaphysical relation that holds between a particular agent and a particular deed, then whenever a responsibility-ascription is correct,

---

[77] This is a more elaborate version of essentially the same distinction, widely used in Dennett's many works. See especially his *The Intentional Stance* (1990).

this is so because it corresponds to such a fact. This would be an epistemological reading of the adverb. Alternatively, "correctly" may just mean "fairly". Correctness as fairness, in relation to responsibility ascriptions, is best tested in situations when responsibility for a particular act confers moral blame and maybe punishment. In such situations we can consult our moral intuitions whether we find it fair to hold such-and-such an actor blameable for such-and-such an act under such-and-such circumstances. This would be a moral reading of "correctly". The two readings may be connected. If there are moral facts, and responsibility is such a fact, then, arguably, it is not fair to ascribe responsibility to an agent for an act unless one knows, or is at least epistemically justified in taking it very likely, that the responsibility relation holds between the action and the agent.

A possible way of arguing for the irrelevance of the epistemic reading: the alleged irrelevance of objective facts about responsibility for our responsibility ascribing practices

Dennett says clearly he is not aiming at correctness in the first reading. Interestingly, he does not say that there is no such thing as responsibility as a metaphysical fact of the matter. He declares it uninteresting:

> Why would anyone care whether or not he had the property of responsibility (for some particular deed, or in general)? Of course people can want just about anything, and yearning for responsibility might arise when one was in the mood for satisfying a purely metaphysical hankering. (Imagine someone who managed to work himself into the state of contracting a desire to eat a piece of bread composed of molecules all of which had once been part of a piece of bread eaten by Alexander the Great. Now imagine someone who managed to affect a yearning for metaphysical responsibility – whatever that is.)[78]

I take it that by this Dennett means that the correct reading of "correctly" in his Claim 4 should be the moral reading, and that whether or not an agent is objectively (metaphysically) responsible for

[78] 1984a, p. 163.

an action has no bearing on whether it is fair to hold him responsible. This claim strikes me as one badly needing support from further argument.

Although the two points are not explicitly connected in the text, on the next page (164) Dennett makes a further claim that could be used as a ground for such an argument. He says that the metaphysical matter of fact about responsibility, if there is one, is completely inaccessible to us epistemically. Now, if we add a further premise, a quite uncontroversial one, namely, that, as our moral practices suggest, we do distinguish between fair and unfair ascriptions of responsibility, and we do it routinely, we seem to have an argument in support of the irrelevance claim:

1 We don't have epistemic access to objective (metaphysical) responsibility.
2 We routinely distinguish between fair and unfair ascriptions of responsibility.
3 (Therefore) the distinction we routinely draw between fair and unfair responsibility-ascriptions cannot be based on the objective fact about responsibility (as it is inaccessible). So it must be based on something else.
4 (Therefore) the objective metaphysical fact about responsibility is irrelevant to our responsibility ascribing practices.

How can we support premise 1? Why think that the moral fact about (even our own) responsibility is inaccessible? Dennett doesn't say explicitly why he thinks so in either of the two books. But he says one thing (which has already been cited) in Chapter 4 Section 1 of *Elbow Room* ("The Problem of the Disappearing Self"), which can be relevant:

> We have to wait and see how we are going to decide something, and when we do decide, our decision bubbles up to consciousness from we not know where. We do not witness it being *made*; we witness its *arrival*.[79]

He seems straightforwardly right: we cannot introspectively put our finger on the event of the decision being made in a way which would inform us on how it precisely comes about. Does it bear the

[79] 1984a, p. 78. Stress in the original.

consequence that we should not know whether we were truly objectively responsible for the decision? (Dennett doesn't say that it does. I am just guessing that this may be his reason in support of the inaccessibility claim.)

Well, it does bear the consequence that we do not have epistemic access to the causal genealogy of the decision. That much is certainly true. If objective responsibility turns on the causal details of the decision's coming about, then it is true that we don't have epistemic access to it. We do not know whether we are responsible for any particular choice because, by introspection, we cannot rule out the possibility that the decision was caused by prior causal factors beyond our control, or popped up just at random, and thereby was not truly our making and so we are not responsible for it. If the objective metaphysical fact about our own responsibility is inaccessible to us, so is the fact about the responsibility of others.

It is not inconsistent with the rest of what Dennett is saying to assume that this is why he thinks we cannot have access to the metaphysical fact about responsibility. He claims that the objective metaphysical sense of responsibility (which he often mocks as "cosmic", or "absolute") is absolutely uninteresting. So it is not interesting either if determinedness (or randomness) undermines responsibility in this sense. It doesn't affect his thesis about the compatibility of the *interesting sense* of responsibility with both determinism and indeterminism.

But he might, of course, have other reasons to think that our epistemic access to objective facts about responsibility is blocked. He does not say.

Now what about premise 2? Yes, we distinguish between fair and unfair ascriptions of responsibility, and we do it routinely. But "routinely" doesn't mean that we do it without any doubt. If "routinely" would mean "doubtlessly", then the conclusion that we discern fair and unfair ascriptions of responsibility on grounds that have nothing to do with the objective metaphysical fact would follow. But "routinely" doesn't mean the same as "doubtlessly". We are all familiar with cases when the anxiety about the fairness of ascribing responsibility to particular classes of agents does arise. Let us examine some such cases.[80]

---

[80] For a very careful discussion of such cases see Chapter 4 of Klein 1990. Here I follow Klein's discussion.

We condemn the wrongdoers who act out of pure and clearheaded selfishness but excuse those who are proven psychotics or suffer from a brain damage. We are also inclined to excuse brainwashed offenders or emotionally deprived ones who were brought up in environments that did not provide for their emotional well-being, on the assumption that it explains their criminal inclination. Isn't it that we excuse these agents because we think the offence they committed was *a result of a state of mind for which they were not responsible* since it was none of their making? In contrast, we won't excuse the driver who drinks and causes an accident, if he decided freely to drink alcohol. The same applies to the troublemaker who consciously decided to give himself an adrenalin injection knowing that it would make him aggressive. It seems that the distinction between responsible and excusable wrongdoers in these simple cases is drawn on the basis of objective facts about the causation of the offensive behaviour. Saying that we confidently distinguish between fair and unfair ascriptions of responsibility despite of our ignorance of objective metaphysical responsibility is not the best way to account for these practices. It is rather that we draw these distinctions on the ground of some objective facts which we think relevant to objective responsibility, although we do not know all relevant facts. Maybe we correctly think that psychotic, brain damaged etc. wrongdoers are not objectively responsible, because we rightly assume that the origination of their actions is incompatible with objective responsibility. We do not know everything that should be relevant to objective responsibility, but we know enough to confidently excuse them. The fact that we do not know everything may affect our confidence in our judgement about those who we do not excuse. We blame some offenders confidently not because we could exclude the possibility that if we knew much more about the origination of their offences we would find them excusable. We blame them with confidence because we find this possibility very remote. But we don't do that doubtlessly. Hence the anxiety about determinism. Determinism threatens to bring those remote possibilities close, because it may prove all cases of offensive behaviour analogous to those of the psychotics and the brain damaged in one respect, which may be decisive for the question of responsibility: that what they did was a causal consequence of a state of mind for which they are not responsible.

So, I think, what Dennett rightly points out in Chapter 4 of *Elbow Room*, i.e., that we do not have introspective access to how our decisions come about, cannot be the ground for his claims that the objective metaphysical fact about responsibility is (a) completely inaccessible to us epistemically, (b) irrelevant to our responsibility ascribing practices.

Unfortunately, I haven't found anything else in the two books that could serve as grounds for these claims.

Nevertheless, both premises of the above sketched argument for the irrelevance thesis are highly intuitive. But in the case of premise 1, this intuitiveness well might be the result of the fact that introspection falls short of either confirming or excluding that what we decide is, in the end, random, or determined by a state of mind for which we not responsible. And in the case of premise 2 we have seen that "routinely" cannot be substituted by "doubtlessly", and without this substitution the premises do not yield the required conclusion.

One thing is sure, however, regardless of what Dennett's exact reasons might be in support of the claims he is making. Our analysis has shown that from the fact that we have responsibility-ascribing practices that aim at fairness, although we are in general ignorant of responsibility as an objective metaphysical fact of the matter, we are not entitled to draw the conclusion that the objective fact about responsibility (if there is such a thing) is irrelevant to the fairness of ascribing responsibility. Without prejudging that determinism is incompatible with objective responsibility I only wanted to show that the irrelevance hypothesis is not the only, let alone the most credible, way to account for our responsibility-attributing practices given our arguable ignorance of objective moral facts.

Another way of arguing for favouring the moral rather than the epistemic reading of "correctly" in Claim 4: the argument invoking Dennett's substantive theory of morality

It is clear though that the mere fact that we do rely on objective facts, e.g. about the causation of action, when we decide whether or not we find it fair to ascribe responsibility to someone, does not prove that responsibility itself is an objective metaphysical fact. In both of his books Dennett suggests that morality is best viewed not so much as a matter of objective metaphysics but as a practical means of enhancing co-operation. Once the emergence of reason and

communication is explained in terms of evolutionary biology and psychology, the emergence of morality fits credibly in the evolutionary picture. Reason and communication makes it possible for a group of conspecifics to negotiate norms of conduct, which serve as "design principles" for co-operation[81], and invent means such as reward and punishment to keep members to them. For groups that have norms, and have blame and praise to make them effective, these institutions have a survival value, helping them competing with other groups. Groups are selected for this capacity. This is not only a naturalistic explanation for the existence of morality according to Dennett, he thinks it also provides powerful insights into its nature. He says morality is an institution invented by us. Moral norms are subject to discussion. The merits of design principles for living in a society can be assessed and the frame of co-operation may be redesigned. Norms have a function. Their function is the evolutionary success of the group and the individuals belonging to the group. The discussion of moral norms is not an epistemological enterprise. There is no truth to be found out about morality. It is more like weighing benefits against costs, bargaining about matters of individual and group interest, and making a compromise. Even in the best reasoned and discussed norms there is always an ineliminable element of arbitrariness.

It is clear that if we accept this account of morality, we should read the adverb "correctly" in Claim 4 in the moral, rather than in the epistemic sense.

So assuming Dennett's account of morality, the truth of Claim 4 turns on whether it is fair to hold a deterministic agent responsible on this particular account. But if we try to answer this question we run into problems that show the untenability of the account.

Fairness is just another moral term. So, to be coherent, the question whether it is fair to hold deterministic agents responsible should be decided on the ground of norms, which are subject to the general theory of morality that has been put forward. One immediate consequence of this is that the answer to the question will necessarily be group-relative. It is fair, that is, it is morally right, to hold deterministic agents responsible as long as it is approved by the institutionalized moral codes of the relevant group. But which group is the relevant group? It is certainly not simply the "we" of our

---

[81] Dennett 2003, p. 268.

present discussion, which potentially involves anyone willing to join. We are not necessarily subject to (even roughly) the same institutionalization of guilt and blame.

We were also instructed that the norms under which we will decide the question will ineliminably contain some arbitrariness. That was one of the reasons to reject the epistemological reading of "correctly". Will we be able to arrive at a general conclusion about the fairness of the practice of holding deterministic agents blameable on the ground of somewhat arbitrary group-relative norms?

Well, perhaps there is an argument to the effect that the norms of different groups cannot be so different that they would yield different conclusions about the fairness of holding deterministic agents responsible. But relativity and arbitrariness will cause apparently insuperable problems for the account anyway. For fairness is not simply relative to the group in which a particular moral code is effective. Norms of fairness, on Dennett's account, are just up to any subgroup that is strong enough to see to it that the norms they agree upon are effective in the whole group. But then musclepower, after all, is a master argument in moral debates, which is absurd.

It may be true that, on the long run, communities that adopt norms that serve the good, and enjoy the consent, of virtually everyone belonging to the group are evolutionarily more successful than others run by dictatorial elites. Dennett offers many evolutionary considerations that point at this direction in Chapters 5 and 7 of *Freedom Evolves*. I grant that it is *not impossible* that the evolution of morality has a clear direction, and the asymptotical morality, toward which the different and somewhat arbitrary moralities of different groups converge, is unique and free from norms enforced by dictatorial subgroups. If it were so, than Dennett would not need to equate moral norms with the actual codes of co-operation that are in effect in actual groups. He would have the option to say that moral norms are the codes of co-operation that figure in the asymptotical design of social co-operation to which actual designs converge. This way the problems caused by relativity and arbitrariness could be avoided.

But the evolution of morality we have so far witnessed in human history doesn't seem to give much empirical support to such an optimistic extrapolation. The convergence hypothesis is very unfirm, to say the least, and I suppose this is why Dennett doesn't try to appeal to the asymptotical design of co-operation when he instructs

us about how to understand morality. On his account, the moral institutions, the current design principles of living in a society, qualify as actual morality by definition, whether or not they were set by a dictatorial elite, and there is no coherent way to draw their fairness into question.

But even if we would grant that the evolution of the somewhat arbitrary moral institutions of different groups is convergent, and the fairness of practices (such as ascribing responsibility to deterministic agents) would be judged by reference to the norms of the asymptotical morality, that would amount to the reintroduction of the epistemic reading, so it would not help Dennett's case anyway.

I do not see how the highly counterintuitive consequence of the account that whatever is enforced by a strong enough subgroup as the actual design of social co-operation *is* morality could be avoided. Once we start distinguishing between fair and unfair rules and rulers, we are invoking norms that are not the actual effective design principles of co-operation in the particular group in question, and thereby the integrity of the account is broken. For this reason I think it would be wise, from the evolutionary theorist's part, to be more modest than Dennett, and claim only to have explained how we might have developed the capacity of entertaining, exchanging, and acting upon thoughts with moral contents, and not also what morality is. It might be the case that we developed a sense of morality through the exercise of inventing rules to make co-operation more effective. But it seems that after having been selected for this capacity for a long time, and having this sense fully developed, we ended up endorsing norms whose legitimacy we do not derive from either their being the actual rules institutionalized in the group we belong to, or directly from their collective survival value. Indeed, our sense of morality makes us critical towards some rules that are both actually institutionalized and have a survival value.

My purpose here is not to argue that moral norms have a Platonic reality. For all that has been said, they may be just our creations. Maybe subject to change and some variation from one group to another. But even if this is what they are, I think it is clear that they somehow achieved autonomy from the evolutionary functions their likely predecessors, the design principles of living together, had.

So, on the whole, I do not doubt that on Dennett's account of morality holding deterministic agents responsible can be fair. But I do doubt that his account captures what we normally mean by morality.

Granting the moral reading of "correctly" in Claim 4, and the proposal of discussing Claim 4 on this reading

I have reviewed two possible lines of reasoning to the effect that the "correctly" in Claim 4 should be read morally rather than epistemically, one that was based on the apparent inaccessibility of objective metaphysical facts about responsibility, and the other that invoked Dennett's substantive account of morality, which is compatible only with the moral reading. I concluded that the first of the two failed because dropping the epistemic reading was not the only, and not even the most plausible, way to interpret our retributive practices in the arguable shortage of access to objective moral facts. The second one failed because Dennett's substantive theory of morality yielded highly counterintuitive consequences, which didn't seem to be avoidable, or certainly not in a way that would preserve the account's being uniquely compatible with the moral reading.

I haven't found any reasons in Dennett's two books for favouring the moral, and dropping the epistemic reading of "correctly" in Claim 4 that would not rely either on the claim that objective facts about responsibility are inaccessible, or on Dennett's substantive theory of morality. It may be my weakness, but I cannot even think of any more arguments that could be offered on his behalf. So I conclude that the claim that the epistemic reading should be dropped is not sufficiently underpinned by argument.

I do believe that responsibility is an objective matter of fact, and therefore "correctly" in Claim 4 should be read epistemically. But I don't want to push this issue any further here. I propose that, for the sake of the discussion, we grant Dennett the moral reading, and check whether our moral intuitions support the fairness of ascribing responsibility to deterministic agents. First I will review Dennett's arguments for the fairness of such responsibility ascriptions, which are not dependent on his substantive account of morality. Then I will give my arguments against.

The argument from social utility

On pp. 159-60 of *Elbow Room* Dennett points to the fact that holding people responsible is a useful institution that helps minimize some sorts of harm in the society. Surely this is true, and this is true

whether or not the agents to be held responsible are deterministic. It is true at least as long as we assume the truth of Claims 1-3, which are necessary for there being any point in setting up an institution whose function is to deter potential offenders from offensive behaviour. It is more than just reiterating the claim from Dennett's substantive theory about the link between morality and utility. It might have a direct appeal, without the substantive theory of morality in the background, to our moral intuitions. At least in case the whole society is composed of deterministic agents we seem to have a sensible moral justification for the practice of holding them responsible. If some (rather plausible) assumptions are made on the number of potential wrongdoers who are prevented from causing harm by the institution of responsibility, and on the measure of harm they would cause, one can safely conclude that almost everybody would be worse off had this institution been abandoned or never introduced. But this is not precisely the kind of moral justification we want. For it is natural to expect that whether or not it is fair to hold a particular agent responsible for a particular deed should depend on what the agent did, his condition at the time of his act, maybe his whole history, but we do not expect it to depend on facts about other agents who are not connected to the deed in question in any way. Suppose that the agent whose moral responsibility is drawn into question is the only deterministic agent among billions of indeterministic ones. If there is a worry about the fairness of holding deterministic agents responsible (and we have seen that there is, when we discussed the case of those who we are clearly willing to exculpate because what they did was a result of a state of mind for which they were not responsible, e.g. the psychotic and the brainwashed), then the argument that the society would fall into a morally regrettable state if we did not hold *anyone* responsible would not speak to this worry at all.

It might be objected that my example is dependent on the assumption that indeterministic agents have a better claim on moral responsibility than deterministic ones, so it begs the question in favour of incompatibilism. Moreover, as one might also object, in the end we are interested in the moral accountability of humans, and, presumably, if one human agent is deterministic, so are all the others, so we will never get a mixed pool of agents in terms of determinedness v. indeterminedness. If it is morally right to hold deterministic agents responsible provided that they are members in a

homogeneous pool, it is *always* morally right to do so (as long as no special responsibility-diminishing circumstances obtain).

Let us consider then only the case of a society made up by deterministic agents. It may be morally justified to hold these agents responsible (under some circumstances) for this practice avoids a lot of harm and suffering at the relatively little price of making the wrongdoers suffer. Still it may be the case that this justifiable practice is just choosing the lesser wrong in a moral dilemma. It may be the case that we do break a norm of fairness we endorse by holding deterministic agents responsible, but by not doing so we would break a weightier moral obligation to their fellow citizens (the rest of the society) that we should adopt practices which further their security and peaceful flourishing and help them avoiding harm. So the moral justification for our responsibility attributing practices Dennett points to does not show that we do not break any norm of fairness we endorse by holding deterministic agents responsible. It only shows that it is morally right as long as the only alternative is to unleash the significantly numerous monsters among us.

But even if we think that it is not a problem as long as we have a moral justification for holding deterministic agents responsible, this account of the moral justification must fail. It is a straightforward consequence of this account of justification that the distinction between wrongdoers we want to blame and those who we are willing to excuse on the basis of facts about the causal origination of their offence we can draw by a threshold, which we set arbitrarily, metaphysically speaking, with the purpose of optimizing the social benefit generated by the institution of blame, relative to the cost represented by the suffering of the punished wrongdoers. Dennett explicitly endorses this consequence (1984a:159-61).[82] But this consequence seems impossible to square with our moral intuitions. For, I think, a change in the practical criteria for responsibility-ascription, which would have a social utility by producing a lower level of aggregate harm at a marginally low additional cost, but which we would nevertheless judge unfair, is perfectly imaginable. I think it is a possibility even if we assume that one relevant component of the cost that may occur in case of such a change is the diminished credibility of the whole institution, as Dennett suggests. This is so because the occurrence of intuitive unfairness is not necessarily linked

[82] 1984a, pp. 159-61.

to an occurrence of loss of credibility (or any other cost). Imagine a case when the unfairness presents itself only to a relatively narrow layer of the sophisticated-in-their-moral-intuitions and, although they find the unfairness grave and in principle they would be able to convince a good number of the less sophisticated, they are unable to generate a public discussion on the issue that would move a considerably large portion of the society, because, say, the heaviness of the political agenda and the ongoing public discussion of a scandalous reality show makes it very difficult for them to attract the attention of the media. I think this simple example shows that the fairness of responsibility-ascriptions is logically independent of their social utility; it is not the latter that justifies the former.

Blame and praise as "the best game in town" for the individual

On pp. 153-4 of *Elbow Room* Dennett argues that the suspicion that the world is deterministic and determinism is incompatible with responsibility we do not take "as the prospect of a welcome holiday" in which we can do whatever we want without running the risk of even feeling bad about it. Being held responsible is something desirable, something we should rationally want. This point is reiterated in *Freedom Evolves* (p. 292) where it is referred to as a social force that opposes the trend of "creeping exculpation", which is a result of growing knowledge of causal factors affecting behaviour, such as genetics, upbringing and environment.

Being held responsible is running the risk of being held blameworthy but also a necessary condition for being subject to praise for our achievements. And not only that. At the expense of agreeing that we should be held responsible for our offences we buy the opportunity to hold others responsible. The benefits of such a deal are obvious. So this is not a bad bargain after all. It seems that "holding people responsible is the best game in town"[83].

It may be true that for most agents it pays off to give their consent to the practice of ascribing responsibility. Their being deterministic does not change this fact. Now, does this entail that it is fair to ascribe responsibility to a deterministic agent? I think it does not, for it is simply unfair to hold anyone to the terms of a contract he never signed just because it would have been rational for him to sign it.

---

[83] 1984a, p. 162.

But isn't it fair to hold responsible someone who had taken many advantages in the past of the institution of responsibility by enjoying praise and admiration for his achievements and by being left in peace by potential offenders who were deterred from committing offences against him? It is not only that it would have been rational for him to sign the contract, it is also that he enjoyed the benefits of the terms of the contract as long as they were beneficial for him. Isn't it like signing the contract tacitly? Why would it be unfair to keep on treating him according to the terms of the contract now that he did something to which the contract links unpleasant consequences?

This is a matter on which our intuitions may diverge. Nevertheless I think it is quite clear that we do not want responsibility to depend on the wrongdoer's history of taking benefit from the institution of responsibility thus far. I do not think we would ever blame A and excuse B for the same act only because, say, A is a sculptor who had long enjoyed praise for his artistic achievement while B has nothing comparable to it in his history, all other things being equal. The moral justification for holding one responsible cannot be just that holding people responsible is a "good game", because in that case we would be morally obliged to look into the details of how good it has been so far for the particular agent in question before issuing any judgement, and it is highly counterintuitive to think that we are.

The argument from the enhancement of imperfect deliberators

In *Elbow Room*[84] Dennett argues that given that finite deliberative processes are necessarily imperfect they need to be aided by "corrective feedback forces" such as our responsibility-ascriptions:

> The (entirely unconscious) organization of memory guarantees that only some approximately appropriate subset of relevant points will occur to one in the time available. … Any style…of self-control must buy some efficiencies at the expense of gambles…. Particular instances of conscious problem solving or decision making must include a somewhat arbitrary decision (conscious or not) to terminate deliberation about the main decision while knowing full well that there still are uncanvassed

[84] On pp. 164-5.

> relevant considerations. … [I]t is an inevitable feature of human character, even perfected to its limit. Original Sin, naturalized. It is wise, however, to adopt policies that minimize the bad effects of these inevitable defects of character. … By somewhat arbitrarily holding people responsible for their actions, and making sure they realize that they will be held responsible, we constrain the risk-taking in the design (and redesign) of their characters within tolerable bounds. When in spite of these best measures people get caught in wrong deeds, their gambles…are simply lost and they ought not to object to paying the assigned penalty.

The problem again is with the nature of this "ought". I guess it is supposed to indicate that the agent is morally obliged to accept the penalty, because it is fairly assigned to him. Now, does the fact that he is necessarily an imperfect deliberator explain why it is fair to ascribe responsibility to him for the outcome of his deliberation? I think our natural intuition suggests that, to the contrary, imperfectness as a deliberator is not a responsibility-generating but a responsibility-diminishing condition. What moral consideration could turn this initial intuitive judgement around?

Enhancing imperfect deliberators by holding them responsible is good for the society and good for the deliberators (assuming that it is good to be a better deliberator). It seems that if it is fair (morally justified) to hold them responsible, it is because of the good generated by this practice. The appeal to the good generated by the enhancement of deliberators by holding them responsible can be split into two sub-arguments, the sub-argument from the good generated for the society and the sub-argument from the good generated for the deliberator. These sub-arguments will be analogous to the two arguments that have just been considered and rejected.

I conclude that these Dennettian arguments show that holding deterministic agents responsible have a survival value for both the group and the individual if we are deterministic, but fail to show that it is fair.

But why shouldn't it be fair to hold deterministic agents responsible?

The argument against the fairness of blaming deterministic wrongdoers: our moral intuitions support an ultimate-origination-condition for moral responsibility

There are two very intuitive conditions for moral responsibility that do not seem to be reconcilable with determinism.

I think most of us would say, before engaging in philosophy, that no agent can be justly blamed for carrying out the best course of action available to him at the time, even if that course of action in itself is not very attractive. If only one course of action is available to the agent, then that is the best. So the agent is not blameable for performing it. The availability of at least one alternative possible course of action with a better moral evaluation is a precondition for blameworthiness. A condition that a deterministic agent can never meet. This condition is often called the Principle of Alternate Possibilities (P.A.P. – Frankfurt 1969, AP principle – Kane 1996), or the Could-have-done-otherwise Condition (C-condition – Martha Klein 1990).

Similarly intuitive is the claim that no agent can be justly blamed for an action that was necessitated by a set of jointly sufficient causal conditions for which he was not responsible. That the causal chain goes through the agent's deliberative faculty makes no difference. If his action was a necessary result of a choice, and the choice was a necessary result of a psychological state (including beliefs and desires) of which the agent was not responsible, then he is not responsible for the action. This condition, also one that cannot be met by a deterministic agent, is often called the Ultimate Responsibility Condition (UR principle – Kane 1996) or simply U-condition (Klein 1990). (In the case of both conditions I will stick to Klein's shorthand.)

The C-condition is a widely discussed one. Compatibilist philosophers either debate that the C-condition cannot be met under determinism, or debate that our moral intuitions really endorse a C-condition for responsibility.

The C-condition for responsibility is the same as the C-condition for freedom. The consequence argument which we discussed in the second chapter is purported to show its incompatibility with determinism. Compatibilist attempts to find a loophole in the argument have been given careful consideration in the second chapter. We concluded that the argument's conclusion that responsibility-entailing freedom, if there is a C-condition for it, is

incompatible with determinism can only be resisted by reinterpreting what we mean by that one "could have done otherwise" (with the conditional analysis of "could", or some other way) so that it does not require the objective existence of alternatives. Another option to uphold the view that freedom is compatible with determinism is straightforwardly denying that there is C-condition for freedom and moral responsibility. Either way, the compatibility thesis is upheld by reinterpreting what is to be compatible with determinism—freedom and responsibility. Indeed, Dennett's theory of morality can be viewed as such a reinterpretation. But, as I have argued above, our moral intuitions seem to oppose this reinterpretation.

But maybe the compatibilist doesn't even have to reinterpret the notion of responsibility to claim that there is no C-condition for it.

One major proponent of the idea that, against all appearances, our moral intuitions do not really support a C-condition for responsibility is Harry Frankfurt. In a highly influential article (*Alternate Possibilities and Moral Responsibility*, 1969) Frankfurt offers a counterexample to the C-condition. In his story an agent, called Jones, commits an offence upon a decision made on his own. However, there is a supernatural psychic manipulator around, called Black, badly wanting Jones to commit his offence, who has the power to see to it that Jones commits it anyway. Had Jones ever been inclined to refraining from it, Black would have intervened. But that never happened. So we have no ground to excuse Jones. Yet, Jones could not have done otherwise.

This counterexample disproves the C-condition as it stands. Yet, even if it is true that Jones could not have acted otherwise, it seems that we hold him responsible because we think there were two types of courses of events available to Jones, one of them leading to his offence through Black's intervention, the other without, these types of courses of events have different moral evaluations, and he opted for a token of the type that is the morally worse. Apparently, we can substitute the C-condition with a C'-condition saying that the availability of alternative fine-grained courses of events that would bear better moral evaluation is a precondition for the agent to be blameable for what actually took place. The C'-condition is just as much incompatible with determinism as the original version was, resists Frankfurt-type counterexamples, and it is strongly supported by normal moral intuitions.

Answering Frankfurt by way of fine-graining the courses of events considered was suggested by Peter van Inwagen. Instead of the original C(PAP)-condition, van Inwagen suggested a PPP-condition for moral responsibility. PPP is the shorthand for the "Principle of Possible Prevention", and the condition essentially is that a person is morally responsible for a state of affairs only if he could have prevented it.[85] This formulation of the condition takes advantage of the fact that states of affairs are practically as finely grained as we want. "Smith's being dead", "Smith's being killed", "Smith's being killed by Jones" are increasingly finely grained states of affairs. "Smith's being killed by Jones through Black's intervention" and "Smith's being killed by Jones on his own" are even more finely grained. Where is the threshold, how finely or coarsely grained a state of affairs need to be for responsibility to be associable with it? I think, naturally, the threshold is at the point where the graining is fine enough for there being at least one morally relevant possible alternative, which was available for the actor whose responsibility was drawn into question. Responsibility should be decided on this level. If there is no such level, then there is no responsibility.

This is problematic as it stands, because if Jones killed Smith on his own, without Black's intervention, then, obviously, he is responsible for the states of affairs that Smith is dead, that Smith was killed, and that Smith was killed by Jones, although none of these could have been prevented by Jones, given Black's intentions and powers. So the condition needs some refinement. The refined condition can be stated as follows: An agent A is responsible for a state of affairs S only if S obtains and A could have prevented it from obtaining, *or* if S is a coarse-grained state of affairs that was realized by a finer-grained state of affairs, and A could have prevented the finer-grained state of affairs from obtaining. With this refinement the condition is equivalent to the C'-condition. (In the C'-condition I added the requirement that the available fine-grained alternative should have a better moral evaluation. I did so, because I wanted to emphasize that, even if there was an alternative, the agent doesn't deserve blame for what he did—the state of affairs he realized—if the alternative was even worse, morally speaking. This requirement can be added here, as well.) Both formulations have their advantages, the

[85] Van Inwagen 1983, Chapter 5.

C'-condition is simpler, van Inwagen's original formulation—preventability—is more directly intuitive.

I find van Inwagen's answer to the Frankfurt-type counterexamples completely satisfactory. But John Martin Fischer complained that it is simply incredible that the presence or absence of such fine-grained alternatives would guide our moral judgements, and that it is incredible that the availability of an alternative in which the agent acts totally unfreely (such as when Jones doesn't kill Smith on his own, but through the coercive intervention of Black) would make an action free in the responsibility-entailing sense:

> The proponent of the idea that regulative control is required for moral responsibility insists that there can be no moral responsibility, if there is but one path leading into the future: to get the crucial kind of control, we must add various alternative possibilities. Now it seems that [he] must claim that the addition of the sort of alternative possibility he has identified would transform a case of lack of responsibility into one of responsibility. But this seems mysterious in the extreme: how can adding an alternative scenario (or perhaps even a set of them) in which Jones does not *freely* [kill Smith] make it true that he actually possesses the sort of control required for him to be morally responsible for [killing Smith]? This might appear to involve a kind of *alchemy*, and it is just as incredible[86]

Now this second part of the objection strikes me as a piece of sheer sophistry. Suppose you are a mayor of a Greek village under German occupation in World War Two.[87] A fanatic SS-officer is in command of the troops that control your village. One day he comes up with the idea that you should give a proof of your dedication to co-operation by executing ten randomly picked citizens of your village. The officer says they are in contact with the resistance movement, but you know, and he knows that you know, that this is a lie. The only alternative, as the officer describes it, is that you will be forced to drink a cup of slow poison, and minutes before you die, when you are already only half conscious and too weak to resist, they

---

[86] Fischer 1995, pp. 141. Italics in the original. (Interestingly, Fischer speaks of voting for Clinton at the 1992 presidential election instead of killing Smith.)

[87] The story resembles one that I read in John Fowles' novel *The Magus*.

will make you do something. He hints that they will give a machine gun in your hand, aim it at the ten victims, and make you pull the trigger. I think it is quite clear that you act totally unfreely, if you choose this alternative, nevertheless the fact that you *have* this alternative seems absolutely relevant to the question whether you are responsible or not. You have the power to bring about an alternative course of events in which you are "not free" but you don't kill anyone willingly, so this course of events may well go with a different moral evaluation, even if the only likely difference in the outcome is that you will be dead, too.

Now, of course, this is not a Frankfurt-type case. If it was a Frankfurt-type case then you wouldn't know that the alternative scenario will also lead to the death of the ten citizens of your village. But it seems "mysterious in the extreme" why would *it alone* render the existence of an alternative scenario irrelevant to the question of responsibility that *you don't know* that it leads to the same result.

Were you, however, in a state of mind, for which you are not responsible—because, say, you were brainwashed, or hypnotized—, that would make it literally impossible for you to resist the officer's will, whether or not he challenges you with canvassing an alternative, then, arguably, you would not be responsible.

Fischer would agree, but he would say that the lack of responsibility in this case is due to the lack of guidance control. It is true that in this latter case both regulative control and guidance control are missing. But Fischer's purpose is to prove that Frankfurt-type scenarios are clear cases of responsibility in the presence of guidance control and in the lack of regulative control, or in the presence of a kind of regulative control which *we know is insufficient to ground responsibility*. And this is not true.

As far as the first part of the objection is concerned, I think it is straightforwardly false to claim that our moral judgements are insensitive to fine-grained differences between courses of events, such as the difference between the two possible ways of Smith's being killed by Jones in Frankfurt's example. Suppose Jones is being trialled for the murder of Smith, and at one point of the trial the defence gives evidence of Black's existence, his superscientific techniques, his desire to see Smith dead, and his documented interest in Jones in the last six weeks before the murder. I think that the court would show interest in the facts concerning Black, but the charge against Jones would not be automatically dropped just because Black was around.

(But, most probably, he would be verdicted "not guilty", if the defence could prove that Black actually intervened.) I think it shows that even if it is true that we normally form judgements on coarse-grained states of affairs such as Smith's being killed by Jones, it is perfectly imaginable that some considerations convince us that we should go finer-grained in order to get a fair judgement. I don't think there would be a natural limit to that.

But even if some exceptionally clever Frankfurt-type counterexamples to the C-condition could be upheld against all incompatibilist answers, Timothy O'Connor argues that, given the very strong intuitive appeal of the C-condition, and its wide applicability in normal responsibility ascribing practices, and given that "Frankfurt cases are extremely contrived and (unless we are badly mistaken about the world) never instanced"[88], the most plausible interpretation of the success of Frankfurtian counterexamples would be that our ordinary thinking misidentified the necessary condition for responsibility by conflating the C-condition with the true condition, which nevertheless the C-condition closely tracks. So even if we abandon the claim that the C-condition is conceptually constitutive to responsibility, we would expect something very much like, and closely connected to, the C-condition to be conceptually constitutive.

However, Dennett in Chapter 6 of *Elbow Room* offers counterexamples of a different type to the C-condition, which cannot be fended off simply by shifting to a C'-condition, plus they have the advantage that they do not involve unlikely characters ("counterfactual interveners") like Black.[89] His two examples are Luther's famous *"Here I stand and I can do no other"* claim and his (Dennett's) own inability to torture innocent persons for a thousand dollars. Dennett's argument is that Luther is not disqualified for the moral praise he deserves for his action because "his conscience made it impossible for him to recant"[90] and, similarly, Dennett is not made into "a sort of zombie programmed always to refuse thousand-dollar bribes"[91] by the fact that his conscience makes it impossible to torture someone.

---

[88] O'Connor 2000, p. 21.
[89] Dennett 1984a, pp. 131-9.
[90] p. 133.
[91] p. 134.

Kane in *The Significance of Free Will* answers Dennett's counterexamples[92]. It may be true that given the psychological state Luther and Dennett were in at the time of the action it was literally impossible for them to do any other than they did, yet, as our normal moral intuitions suggest, what they did was not morally insignificant. They deserve the moral praise for their action. But this is so, according to Kane, because we assume that both Luther and Dennett were responsible for the state of mind they were in at the time of the action. They were praiseworthy for being the man they were at the time of the action. That is why they are praiseworthy for the act that was necessitated by their psychological state (by their being who they were at the time of the action). Had the necessitating psychological state been something for which they were not responsible, both the mental state and the action would have been morally neutral. So the defence of the C-condition against Dennett's arguments is based on the U-condition.[93]

The relation between the U-condition and the C-condition is the following. The U-condition does not require that every action for which we are to ascribe blame or praise would be such that the agent could have done otherwise, but it does require that however we specify a set of causal conditions that are jointly sufficient for the obtaining of actions that confer praise or blame would contain an action that the agent could have done otherwise. So there is no U-condition satisfying action without a C-condition satisfying action in its causal history.

Perhaps we are willing to praise one for one's character even if he is not responsible for it, like in the cases of Luther and Dennett in the above examples because it makes no harm. But it seems plain that we are much more cautious to ascribe blame for offences stemming from a character which the agent could not help having. Martha Klein in Chapter 4 of her *Determinism, Blameworthiness and Deprivation* (1990)[94]

---

[92] Kane 1996, pp 38-40, 78-9.

[93] Martha Klein argues at length in her 1990 book that the C-condition, in the sense of it that can be upheld against compatibilist criticism, is not independent of the U-condition, and that the U-condition itself gives enough ground for the libertarian "to be anxious about determinism".

[94] I will not restate every detail of Klein's argument to the effect that our moral intuitions do support a U-condition for blameworthiness, for that would be practically reiterating the whole fourth chapter of her book. I find her arguments both very cautious and highly satisfactory, and I have nothing to say about this particular issue what she hasn't said. So I only give an overview of her argument. For further details please refer to her book.

analyses such cases and draws the conclusion that, though not all of us are explicitly committed to a U-condition,

> some of our moral intuitions are 'U-condition generating beliefs', that is beliefs which commit those who hold them (whether they realize it or not) to the belief in a U-condition. These intuitions are all beliefs to the effect that agents are not morally responsible if their actions are caused by certain specific factors; what these factors have in common is that they are states or events for which the agents are not responsible.[95]

Among these intuitions there are beliefs about offenders whose offences are the outcome of a state of mind which is attributable to brain damage or brainwashing, to the effect that it would be unjust to hold them responsible. Klein argues that it is the fact that they are not responsible for the cause of their offensive behaviour, rather than any other fact, that accounts for our willingness to excuse them.[96] We have similar attitudes toward emotionally deprived wrongdoers. Before passing judgements on them we naturally ask ourselves whether they can be held responsible for having the criminal inclinations they have.

> Information…about the childhoods and upbringing of young delinquents is increasingly taken into account by magistrates, police, social workers, before recommendations and orders are made in respect of them. Often offenders who have been deprived of affection are thought to be not so much candidates for punishments as candidates for an environment which will help to make up for what they lacked.[97]

Klein considers other candidate explanations, on the compatibilist's behalf, for our having these attitudes toward such cases and rejects them before she arrives at the conclusion that the

---

[95] Klein 1990, p. 66.
[96] p. 67.
[97] p. 69.

only satisfactory explanation for our reactions to these problem cases is that we are implicitly committed to a U-condition.[98]

Now, if our moral intuitions endorse a U-condition for responsibility (at least for blameworthiness), then these very same intuitions disapprove holding deterministic agents responsible, for a U-condition is clearly incompatible with determinism.

It seems to be a strong enough ground to think that holding deterministic agents responsible is unfair, regardless of the fate of the C-condition, that is, whether or not our moral intuitions really endorse it, or, whether or not it is really incompatible with determinism.

Summary

I have stated my reasons to reject Dennett's Claim 4. These reasons were: a) that as far as a positive account of the nature of morality can be read into the two books (especially *Freedom Evolves* invites such a reading), which would guarantee that deterministic agents can be properly said to be morally responsible, I find this account impossible to square with our moral intuitions; b) that the arguments Dennett offers for the moral justification of holding deterministic agents responsible (mainly in Chapter 7 of *Elbow Room*), independently of his positive account of morality, I do not find convincing; and c) that may or may not the arguments Dennett offers (throughout both books) against a C-condition for moral accountability deserve some merit, our moral attitudes seem strongly support a U-condition for the same, making a strong case against the fairness of blaming deterministic agents.

This much in the name of Conrad.

As far as our general project is concerned, the following conclusions are in line. The above arguments seem to establish that our moral intuitions support a U-condition for blameworthiness. The U-condition requires that there be self-forming actions. Self-forming actions are such that the agent could have done otherwise, not only in the "practical perspective" (e.g. in the sense of the conditional analysis, or only adopting the "personal stance"), but in the "theoretical perspective", as well. So they are incompatible with determinism—this incompatibility claim only requires the

[98] pp. 69-75.

consequence argument to go through until its intermediary conclusion 4 (that we cannot make any difference to the present or the future, if determinism holds). So compatibilist reinterpretations of freedom, which render freedom so that it doesn't require that at least sometimes there be alternative courses of action, or only in the "practical perspective", are not strong enough senses of freedom to ground responsibility, in the sense required by our moral intuitions. But as the discussion of the consequence argument in the second chapter has shown, so reinterpreting freedom is the only way available to the determinist to maintain the thesis of the compatibility of freedom with determinism. From this we may conclude that no compatibilist senses of freedom can be strong enough to ground moral responsibility, in a sense required by our moral intuitions. So, in this respect, libertarian freedom, provided that it is possible, is better than compatibilist freedom, for libertarian free actions meet the U-condition. And this is what I wanted to establish in this chapter.

Further worries

Further questions need to be answered, however. There is a due worry that the U-condition might be incoherent. If it is coherent, there still might be reasons to think that self-forming actions, which must figure in the causal ancestry of any action for which we can claim responsibility on the U-condition, cannot occur out of rational choice (which would limit the scope of genuine freedom to irrationality, and make it very unlikely that a U-condition is indeed a condition for responsibility), or that they are impossible to detect reliably (which would deprive the U-condition from any practical applicability in our moral practices). I will address these questions in due course, once I am finished with the question whether we need and can have genuine alternatives.

5 Do We Need Genuine Alternatives? – B. An Epicurean Meditation

This chapter is going to be about rationality. I would like to note that not every claim I make in this chapter is supported by arguments I hold to be conclusive. Hence the title of the chapter (meditation). Nevertheless, I think the arguments provide strong support for my main claim that rationality cannot be the property of a thought produced by a mechanism. I will indicate where the points are where I think my arguments are inconclusive.

Our relationship to the truth depends on whether we are free in the libertarian sense or only in the causal (compatibilist) sense. The Epicurean argument

We relate ourselves to the truth by relating ourselves to the truth or falsity of propositions. We *know* they are true or false, or *hope*, or *doubt*, or *believe* them, etc. These relations are called propositional attitudes.

It doesn't just happen to us that we have propositional attitudes. Having them is not always a passion. In many cases we come to propositional attitudes by way of epistemically, and in many cases also normatively, evaluating propositions, forming a judgement on the ground of these evaluations about them, and committing ourselves to certain relations toward them.

This active character of propositional attitudes is a point where determinism may cause problems, because determinism has a ring of passivity.

The Athenian Epicurus, an early champion of the libertarian conception of free will, in the third century before Christ tried to show that determinism is a self-defeating doctrine on this ground. He said that the determinist theorist cannot really criticize the indeterminist theory, and cannot really argue for his own view, because his view is that whatever he holds true of the causal organization of the world, or of whatever else for that matter, and whatever arguments he entertains in support of it, are just effects of

causal factors of which he is a helpless subject.[99] Determinism claims of itself that it came into being the wrong way, namely, passively.

Intuitively, the Epicurean concern seems well founded in at least that a deterministic cognizer cannot know, or cannot be said to have related himself to a proposition in any way that requires epistemic (or normative) evaluation, judgement and commitment in quite the same sense as a non-deterministic cognizer can. No doubt, a deterministic cognizer can go through a process, which, for all its introspective phenomenological features, feels like a process of epistemic evaluation. It involves considering and evaluating the reasons for and against taking something true, and actually making the commitment that he relates himself to the proposition the given way. But it is far from clear that it is of the same value, or that it counts as epistemic evaluation and commitment in the same sense, as in the case of a non-deterministic cognizer, having regard to the fact that it was never really possible for the deterministic cognizer to consider other evidence than he did, evaluate the ones he actually considered differently, or refrain from making the commitment.

Perhaps there would be nothing wrong with being caused to take a proposition to be true any time when the proposition in question is the logical consequence of evidently true premises, or premises we already wholeheartedly endorse. Perhaps we should not yearn for freedom from a built-in machine-like logical calculus, a logical engine, if we had a thing like that. (Nevertheless, unluckily or not, if it exists, we seem to have freedom from it, as it is made evident by the many logical mistakes we commit.) But most propositions to which we relate ourselves epistemically, one way or the other, are not of that sort. Since in most cases the reasons actually considered do not seem to absolutely necessitate the resulting epistemic attitude, there is something worrying in the thought that the resulting attitude is nevertheless caused. The worry is that then it might be the result of something else, not the judgements made on the evaluation of reasons.

The Epicurean argument does not refute determinism, but that is not the point

A present day advocate of determinism, Ted Honderich argues that the objection does not refute determinism. There is no

[99] See Epicurus Fr. 34, 26-30 (*De natura*) or *Sententiae Vaticanae* 40 in von der Mühl 1922.

contradiction, he says, between a theory being produced the causal way and it's corresponding to a fact.[100] This is true, but beside the point. The point is that the determinist theorist cannot claim that he knows his theory is true. Because knowing (in the sense that is now applicable) is an active propositional attitude and he takes himself to be deterministic and, therefore, passive. Determinism seems to lump his putative knowledge with the beliefs that are produced automatically in him, like, for example, a belief about what the weather is like is produced automatically in a cognizer who is appropriately positioned for such a belief to be perceptually produced in him.

Suppose there is a causal machinery in the determinist theorist's head that, in response to environmental stimulation, produces propositions that track the truth very reliably. It is nice to have such a machinery, but, and this seems to be Epicurus' worry in modern terms, the question is not whether he can think such propositions with assent, and whether they can track the truth, but whether *he is related to these propositions* in a way a knower should be related to what is known.

On the face of it, the deterministic cognizer's relation to the facts represented by these propositions is something like the thermometer's relation to the temperature.[101] A thermometer easily beats us in telling whether it is warm outside or not. But it does not *know* the temperature.

Two possible objections to the Epicurean conclusion that the determinist cannot know that determinism is true

At this point, a determinist could object two ways. He may complain that what I have said about activity and passivity in relation

---

[100] Honderich 2002, pp. 88-90.

[101] D. M. Armstrong, one of the first proponents of the theory that knowledge is a belief that is caused in a certain way, so that it tracks the truth, himself calls his position "the thermometer model of knowledge". (Cf. Armstrong 1973, pp. 162-83, reprinted in Bernecker-Dretske 2000, pp. 72-85.) One important difference between Armstrong and Alvin Goldman, the other founding father of reliablism, is that whereas Goldman presented his reliablism as a third-personal description of what is the case when justification, an essentially first-personal phenomenon, obtains (cf. *"The justificatory status of a belief is a function of the reliability of the processes that cause it, where (as a first approximation) reliability consists in the tendency of a process to produce beliefs that are true rather than false."* (Goldman, 1979, p. 10)), Armstrong claimed that justification, the first personal fact, was altogether unimportant for knowledge. We will come back to this difference later.

to knowledge is dependent on a very old and outmoded idea of what knowledge is, and that he knows better and more up to date ones, which do not require the kind of activity I have depicted. Alternatively, he may simply protest against taking deterministic agents to be passive in the sense that corresponds to the sense in which adopting propositional attitudes is an activity. Let us consider these objections one by one.

Does the Epicurean conclusion depend on an outmoded account of what knowledge is?

The suggestion that knowledge is true belief which we have good reasons to embrace, i.e., we believe because we are *epistemically justified* to believe, is very old. It is older than the Epicurean objection to determinism. It is there already in Plato's *Theaetetus*.

The incompatibility between knowing and being deterministic, if it arises, arises because of justification. The problem is that it is unclear if a deterministic cognizer can be said to have a belief as a result of an epistemic evaluation of the required kind. It has already been acknowledged that a deterministic cognizer can go through a process which, for all its phenomenological features, is like a process of epistemic justification. It involves considering and evaluating the reasons for and against taking something true, and actually making the commitment that he should relate himself to the proposition the given way. But the question is whether this really counts as a justificatory process if, among other things, it was never really possible for the agent to consider other evidence than he did, evaluate the ones he actually did differently, and refrain from making the commitment. Surely, the whole issue would not arise, if justification was not required for knowledge.

The view that knowledge is justified true belief enjoyed near hegemony until 1963 when Edmund Gettier in a three pages long paper produced two counterexamples to the justified true belief analysis of knowledge.[102] The Gettier counterexamples are widely considered to have sunk that account of knowledge, once and for all. Here is one of them in his own words.

[102] Gettier 1963.

> Let us suppose that Smith and Jones have applied for a certain job. And suppose that Smith has strong evidence for the following conjunctive proposition:
>
> (d) Jones is the man who will get the job, and Jones has ten coins in his pocket.
>
> Smith's evidence for (d) might be that the president of the company assured him that Jones would in the end be selected, and that he, Smith, had counted the coins in Jones's pocket ten minutes ago. Proposition (d) entails:
>
> (e) The man who will get the job has ten coins in his pocket.
>
> Let us suppose that Smith sees the entailment from (d) to (e) and accepts (e) on the grounds of (d), for which he has strong evidence. In this case, Smith is clearly justified in believing that (e) is true.
>
> But imagine further, that unknown to Smith, he himself, not Jones, will get the job. And, also, unknown to Smith, he himself has ten coins in his pocket. Proposition (e) is true, though proposition (d), from which Smith inferred (e) is false. In our example, then, all of the following are true: *(i)* (e) is true, *(ii)* Smith believes that (e) is true, and *(iii)* Smith is justified in believing that (e) is true. But it is equally clear that Smith does not *know* that (e) is true; for (e) is true in virtue of the number of coins in Smith's pocket, while Smith does not know how many coins are in Smith's pocket, and bases his belief in (e) on a count of the coins in Jones's pocket, whom he falsely believes to be the man who will get the job.

What Gettier has shown is that conditions (i) – (iii), i.e., that the proposition in question is true, is believed, and is believed justifiedly, are not jointly sufficient for the proposition to be known. His counterexamples are dependent on two assumptions, as he himself notes. One is that one can be justified to believe a proposition which is actually false, the other is that if one is justified in believing a proposition then he is also justified in believing any further propositions that are logically entailed by the first one, provided that he sees the entailment. Justifiedness is transferred through seen logical entailment, that much, I think, should be granted to Gettier. As far as the first assumption is concerned, I think that is not so

plainly uncontroversial, but it is certainly true that if we set the standards for justifiedness so high that they warrant against getting the thing wrong, then most cases of what we normally call knowledge would not qualify as knowledge on the justified true belief account of knowledge. The way Smith is justified in the above example in believing that Jones will get the job and has ten coins in his pocket is not worse than the level of our justifiedness in believing many propositions that we normally take to be known. Requiring that justification should be truth-warranting would be a departure from the way the term "knowledge" is normally used.

Gettier's conclusion that conditions (i) - (iii) fail to be jointly sufficient for knowledge is particularly important for someone who wants to know what knowledge is, since an analysis would require the identification of conditions that are both necessary and sufficient. It is important to note that the example does not show that any of these conditions would not be necessary. For the problem for determinism to arise it is enough if justification is a necessary condition for knowledge. Our present concern is not the correct analysis of knowledge. So, on the first look, it seems that we can just neglect the Gettier examples. On second thought, however, it seems also true that our reason to think that justification was a necessary condition for knowledge was that we thought we knew what knowledge was. It suggests that Gettier's conclusion does concern us, after all.

Now, how exactly Gettier brought down the justified true belief analysis?

As we have already noted, the counterexample was dependent on the assumption that there may be a gap between the truth of a proposition that is justifiedly believed and its justification. In fact all Gettier-type examples are dependent on this assumption, and any attempt to improve on the justified true belief analysis of knowledge by introducing additional conditions is bound to fail unless the additional conditions close that gap, because as long as the gap is there, however small it is, the account is vulnerable to Gettier-type counterexamples.[103]

---

[103] Suppose Smith is justified in believing proposition p. If the gap is there, the proposition can be false. Suppose it is false. Suppose p entails q, and Smith is aware of that. Since justifiedness is transferred through seen logical entailment, Smith is justified to believe in q. Suppose that q is true, and it is true in virtue of pure luck, i.e., in virtue of things that have nothing to do with Smith's justification for p. This is the recipe for producing Gettier counterexamples. As long as the gap is there this recipe works, as it was pointed out by Zagzebski (1999).

Now, not everybody agrees that a gap should be allowed between the standards of justifiedness that we set for knowledge and truth. Descartes' motivation in his pursuit of the Method was exactly that he wanted the gap closed in order to escape scepticism. In the *Republic* Plato seems to distinguish knowledge from other epistemic states, prominently true opinion, partly on this ground, too. His view is that in the case of knowledge no such gap is allowed. Surely, if the standards of justification are as high as, for example, in the case of Cartesian foundationalism, according to which we are considered to be justified to believe in *p* only if we see that *p* is logically entailed by propositions that are absolutely inconceivable to be false, then there is no room for Gettier counterexamples.

In order to distinguish justifiedness in this truth-warranting sense from normal justifiedness, let us write the former with a capital J. Although, as it was already admitted, it would be a departure from the way the word is used in everyday contexts, it would not be unnatural to agree on the terminological convention that we should call knowledge only Justified belief. Truth does not need to be mentioned as an independent third condition since Justifiedness entails truth. When justification falls short of being truth-warranting but is still good enough by some standards, we could call the corresponding propositional attitude "justified belief", whether or not the belief is true. It can happen to be false if we are unlucky enough, because of the gap that there is between justification and truth.

But it is not nice to use words differently from their normal use, unless it is necessary for avoiding confusion. Now it is not absolutely necessary. It is admissible to call some of the justified beliefs knowledge. The numerous lucky ones, which happen to be true. Confusion may be avoided by using lower case and capital k's. We may agree on the convention that from now on we will write the first kind of knowledge, the one that goes with Justification, with a capital K, and with a lower case k the second kind, which is knowledge in virtue of justification (with lower case j) plus some luck.

Surely, it is legitimate to distinguish between Knowledge and knowledge, since the epistemic states of the Knower and the knower are not the same. They are not related to the respective propositions the same way. If Smith was man of reflection, he should have had at least some doubt about his belief that he would lose the competition to Jones for the job. It was improbable but not impossible that he would win after all. Surely, he should have been less confident in his

judgement about this proposition than Descartes was about his own existence, for example.

Equally natural it is to distinguish between knowledge and the other instances of justified belief. But not in the same respect. The facts that distinguish knowledge from other instances of justified belief are external to what the subject actually accesses in his consciousness when he forms his judgement about the propositions. Were they internal, they would figure in the justification and turn it into Justification. If we allow for this talk of knowledge, with lower case k, we are bound to be "externalists" about knowledge at least in this sense. (Although actual externalists hold the stronger thesis that the facts that are critical for knowledge can be external also to whatever is *potentially* accessible to the knower within his consciousness. If it can be external to what is actually accessed, this further step does not make too much of a difference to what knowledge is like from the internal, first personal perspective.) Note that there is no contradiction between being an externalist about knowledge and an internalist about Knowledge. Surely, knowledge and the other instances of justified belief differ in that the former relates us to propositions that are true and the latter don't—but they feel the same from the inside. So even if it is true that they are distinguishable in a principled way, it doesn't follow that they fall into different classes of *propositional attitudes*. Not on our definition of propositional attitudes, on which they are ways of relating ourselves to propositions. What distinguishes between instances of knowledge and other justified beliefs is something that is fully external to the activity of relating ourselves to the proposition in question, i.e., the process of evaluation, judgement and commitment. One way of putting it is that the difference does occur in a third-personal description of our epistemic status, but does not occur in the first-personal perspective. So, as propositional attitudes, they should not be classified into different categories. Suppose we learn that God really exists, but God himself forbids us to tell it to anyone. Would we think that so far, before we learned that the proposition was actually true, we somehow misrepresented the attitude the believers bore to the proposition that God exists? Would the fact that He does, unknown (unKnown) to anyone but us, prove that their belief in Him is in fact a different propositional *attitude* from the one we so far thought it was? I don't think so. The distinction between knowledge and other instances of justified belief is of the same kind. This

difference doesn't make them belonging to different types of propositional attitudes. It is more appropriate to view them as different instances of the same type of propositional attitude, which differ only in how lucky they are.

We are not interested in how lucky the determinist theorist is in thinking that determinism is true. So it seems that, provided that we agree upon the proposed convention of calling lucky justified beliefs knowledge (lower case j, lower case k), what we are interested in is not whether he knows that determinism is true, but whether he is justified in believing that determinism is true. At least. If he is Justified, not just justified, that is also a relevant fact. But whether just justified, or justified plus lucky, that is not a relevant distinction, as long as we are exploring his relatedness to the thesis of determinism, because facts external to what is accessible to his consciousness cannot influence it and cannot have a bearing on the question whether determinism can be claimed in an intellectually responsible way.

But if it is accepted that knowledge (with lower case k, meaning justifiedness plus epistemic luck) is not an interesting notion for our present purposes, we can go one step further, and say that any other version of knowledge (italicized *knowledge*, for example), which the externalist epistemologist[104] may propose, one, for example, that doesn't require justifiedness at all, but which is defined in terms of purely external facts (external to what is accessible to the subjects consciousness) that don't necessarily covary with internal facts about justification, is also uninteresting for our purposes, for the simple reason that *knowledge* on such an account would not be a kind of propositional attitudes at all. Of course, some, probably most, instances of *knowledge* can be propositional attitudes, but the predicability, or the lack of predicability, of the property *knowledge* does not distinguish between different types of propositional attitudes in our sense, since such distinctions are sensitive exclusively to facts that are internal to the subject's consciousness.

So, in order to separate ourselves fully from the externalism-internalism controversy about knowledge, all we have to reply to the first objection of the determinist is that, since we do not presently want to defend any position concerning the nature of knowledge, we embrace *the possibility* that we might have misstated Epicurus'

---

[104] Armstrong, for example—see footnote 102 earlier.

objection against determinism when we said that the point was that the determinist cannot *know* that determinism is true, and so we restate the conclusion of the argument in terms that concern only how the determinist relates himself to his thesis. That will do for our purposes.

It has already been admitted that, as far as the Epicurean argument is concerned, determinism can be true, and I suggested that our concern should be rather whether the determinist can know it. Now, for the sake of escaping further duties in epistemology, I admit the possibility that on some respectable accounts of knowledge the determinist may know that determinism is true, and shift my focus to the question whether he can have a justification for believing in determinism that is fully accessible to his consciousness. One way of stating that he can is saying that the truth of determinism is transparent to him. Another way of stating this is saying that he has an attitude toward the proposition that determinism is true, let's call it **knowledge** (now a fourth kind: bolded knowledge), that has the property that, if he **knows** that determinism is true, then he **knows** that he **knows** that determinism is true.[105]

Why should we require that? Well, if it can be shown that on the determinist hypothesis he cannot know that he knows that determinism is true, then what he says in support of determinism is not said in an intellectually honest and responsible manner, because then he is trying to convince us of a thesis he believes to be true, but he doesn't fully know why, or he knows why, but he doesn't know if this ground of his is a good enough ground to believe in the thesis. Had he known why, and had he known that his ground is good, then he would know that he knows that the thesis is true. So, if it can be shown that he cannot know that he knows, then determinism is either false, or it cannot be advocated in an intellectually honest and responsible way. This is the sense, then, in which, on behalf of Epicurus, I suggest, we may regard the thesis of determinism self-defeating.

---

[105] Second order knowledge (knowledge of knowing) is a warrant of transparency. It is pointless to go beyond the second order. For if the notion of knowledge we are using is such that, because of the nature of the epistemic justification involved in the notion, knowing that p implies knowing that knowing that p ($Kp \rightarrow KKp$, often called the "KK-principle"), it is easy to see that it implies transparency also in the millionth order, or what we have. (Substitute Kp in place of p in the formula stating the KK-principle. And so on.)

Talk of knowledge could have been avoided all the way through. It should be admitted that if the determinist could show that his justification for believing in determinism is transparent to him (he knows he is justified to believe in determinism, or, better, he is justified in believing that he is justified in believing in determinism) this would amount to the fall of our Epicurean objection against determinism, even if, in virtue of a gap between justification and truth, determinism happened to be false. If the determinist could be as much justified to believe in determinism, in a way that is transparent to him, as Smith was justified to believe that Jones would get the job and had ten coins in his pocket, that would be enough trouble for Epicurus.[106]

So much about the objection against the Epicurean argument that it is based on an outmoded conception of knowledge. It isn't. It is based on the simple idea that the determinist to be fit for advocating his thesis in an intellectually honest and responsible manner must be justified in believing it in a way that is transparent to him, whether or not that is also a necessary condition for knowing the thesis.

Was it correct to describe deterministic cognizers as passive in a sense that entails that they cannot be transparently justified in holding propositions true?

Perhaps the best way to approach the question whether it is right to characterize the determinist cognizer as passive in a sense that would pre-empt and invalidate his justifiedness is to give a description of the process through which a cognizer can get justified in believing a thesis of a level of abstraction and generality similar to that of the thesis of determinism, and see where the points are in this process at which deterministic and non-deterministic cognizers may differ so that these differences affect their epistemic states.

A model of the process of justification

(i) First of all, such a process requires the capacity to have mental states that represent states of affairs outside the cognizer's mind.

---

[106] We could use J's instead of K's in the principle in the previous footnote. Even if there is a respectable account of knowledge on which the KK-principle does not hold, it is much harder to imagine a respectable account of justification, on which a JJ-principle does not hold. Another way of saying this is saying that the already familiar "knowledge" (with lower case k) would perform perfectly well in the role of "**knowledge**" (bolded knowledge). And so would, of course, "Knowledge" (with capital K).

(ii) The cognizer should be justified in trusting that some of these representations of particular facts are adequate. This, of course, involves the capacity to discern veridical representations from others.

(iii) He must be capable of having mental states that represent abstract facts. At these he arrives by way of empirical generalizations.

(iv) At the latest at this level it becomes quite obvious that *"to know one must think"*.[107] In order to achieve even a minimal level of generality, the subject matter should be adequately conceptualized, which means that the logical relations of the representations involved should be explored.

(v) In order to see whether his inductive reasoning is trustworthy, the cognizer must be capable of forming a judgement about the weight of the empirical evidence that supports his generalizations. Up to a point he relates himself to every general proposition he is entertaining as one should relate oneself to a hypothesis which is possibly true but can also be mistaken. But then his relation to them changes. Under the weight of evidence his doubt diminishes. When he starts using one of them as a ground for further reasoning, this reflects a commitment to the probable truth of the proposition, at least to the extent that he thinks that any further reasoning on its ground is worth the effort. This usually involves being sensitive to potential counterexamples that would falsify it. From the requirement that his justifiedness should be transparent to him it follows that the cognizer should be confident that he knows where to look for counterexamples and should trust that he would see them if there were any.

(vi) The reasoning that leads to the thesis might involve the drawing of deductive inferences as well. For that it is required that there be norms of deductive inference, which, when followed, guarantee that one doesn't get false conclusions from true premises.

(vii) The cognizer has to see the truth of these norms—the laws of logic.

(viii) He needs to be able to evaluate critically his deductive reasoning, which involves identifying his premises and the deductive steps he is taking, and checking whether the steps are really endorsed by the norms of deductive inference, and that the premises used are ones that he really trusts.

---

[107] Willard 2000.

(ix) In the end, he has to form an overall judgement on his epistemic position relative to his thesis, taking into account the firmness of its empirical basis, and the goodness of both the inductive and the deductive reasoning he invoked in support of it, and decide how he should relate himself to the thesis, after all. In many cases this activity involves the consideration of possible objections to the thesis and a careful screening of the whole reasoning for possible weaknesses or mistakes.

(x) It must be added that a known proposition never stands alone in the consciousness of the cognizer. Through the conceptual relations the representations involved bear to other representations (see point (iv)), and through the logical relations of the proposition with other propositions that were explored in the process of deductive reasoning, a known proposition is always embedded in a larger context of what is known. Knowledge is not discrete; it rather maps larger areas of reality as a whole. This feature of knowledge, following Kant, is sometimes referred to as the "noetic unity".[108] It bears the consequence that the justification of particular propositions has always a context. One important aspect of being justified in holding a proposition true is that the proposition is part of a web of beliefs (to borrow a Quinean term) that is justified as a whole.

I think most actual proponents of the determinist thesis would report that their justification in believing the thesis consists of, or is dependent on, such items. Most of them, I believe, would also report that they have taken some information provided by others, maybe authorities in science or in philosophy, for granted, on the basis that they know that the provider is a very reliable source of information. But, surely, very few would say they believe in determinism just because someone of very high esteem said it was true. Or at least not with the ambition to convince us of determinism. The elements of justification just listed cannot all be substituted by relying on other people. And relying on others is ultimately dependent on believing that they are justified in believing in the thesis by ways described by items (i) – (x). So belief-acquisition by way of relying on an authority is not an interesting case that should be given an independent discussion.

---

[108] Cf. Willard, ibid.

Now, which are the items on the list from (i) to (x) in respect of which the deterministic agent differs from his non-deterministic counterpart?

Well, most of them. The procedure of justification just sketched is a minefield for the determinist theorist. Almost all of the items on the list have been tried to be shown impossible to perform by a determinist cognizer, or just impossible if determinism holds, or if the broader metaphysical theory, in the context of which determinism is usually advocated, i.e., mechanistic materialism, is true.

(Mechanistic materialism is an old term. In newer terms, it is a combination of ontological physicalism (the theory that everything that really is is either physical or supervenes on the physical) and the idea that physics is deterministic and physical causation is *mechanistic*, that is, *devoid of purpose*. There is a thesis that links these two claims together. That is the thesis of the causal closure of physics, as we have seen in chapter 2. So far I haven't emphasized the mechanistic character of physical causation, but now it will be important, because one thing we will discuss is the question whether there is a place for purpose and reason in a world that is at bottom a physical mechanism.)

Intentionality, truth and normativity

Already the first item on the list, on which all the rest are dependent, propositional content (representation, meaning, reference, aboutness, intentionality, no matter how we call it) has been repeatedly argued to have no place in a world which is, at bottom, a material mechanism.

> An ant is crawling on a patch of sand. As it crawls, it traces a line in the sand. By pure chance the line that it traces curves and recrosses itself in such a way that it ends up looking like a recognisable caricature of Winston Churchill. Has the ant traced a picture of Winston Churchill, a picture that *depicts* Churchill?
>
> Most people would say, on little reflection, that it has not. The ant, after all, has never seen Churchill, or even a picture of Churchill, arid it had no intention of depicting Churchill. It simply traced a line (and even *that* was

> unintentional), a line that *we* can 'see as' a picture of Churchill.
>
> We can express this by saying that the line is not 'in itself' a representation of anything rather that anything else. Similarity (of a certain very complicated sort) to the features of Winston Churchill is not sufficient to make something represent or refer to Churchill. Nor is it necessary: in our community the printed shape 'Winston Churchill', the spoken words 'Winston Churchill', and many other things are used to represent Churchill (though not pictorially), while not having the sort of similarity to Churchill that a picture – even a line drawing – has. If *similarity* is not necessary or sufficient to make something represent something else, how can *anything* be necessary or sufficient for this purpose? How on earth can one thing represent (or 'stand for', etc.) a different thing?

These words are, of course, not mine. These are the great Putnam's.[109] I couldn't put it any better, so I let him continue.

> The answer may seem easy. Suppose the ant had seen Winston Churchill, and suppose that it had the intelligence and skill to draw a picture of him. Suppose it produced the caricature *intentionally*. Then the line would have represented Churchill. (…)
>
> But to have the intention that *anything*, even private language (even the words 'Winston Churchill' spoken in my mind and not out loud), should represent Churchill, I must have been able to think about Churchill in the first place. If lines in the sand, noises, etc., cannot 'in themselves' represent anything, then how is it that thought forms can 'in themselves' represent any thing? Or can they? How can a thought reach out and 'grasp' what is external?

It seems that there is a mystery here. It turned out that the ant's drawing has no original intentionality. If it has any intentionality at all,

---

[109] 1981, p. 1. Stress in the original.

it is derived intentionality, derived from the more original intentionality of our thoughts, or the thoughts of the ant, if it thinks. The same is true about all signs we are using, including linguistic signs, like words. So we can account for their intentionality if we account for the intentionality of our thoughts. But how could this be done? Surely, just saying that there is something in our head that is pretty much like a language, a language of thought, say, of which natural languages are just translations, which have elements, like natural languages consist of words, which can be called mental representations, would not do. The linguistic analogy, even if it is true in some important respects, will surely not explain how mental representations are representations, how it is that they are intentional. For the intentionality of words is derived intentionality, derived from the very intentionality of mental representations. Surely, we cannot give an analogous explanation for the intentionality of mental representations themselves.

The mystery can be given a name. *Original intentionality*, thought of as the irreducible characteristic of minds (selves, persons) that in their inner lives they are related to other things in a conceptually and metaphysically primitive way, would be the name. This would serve as the foundation for all cases of one thing standing for another.[110]

Surely, it doesn't explain much. That's why it is mystery. But not every mystery is bad philosophically. If there is nothing that could be explained, then the lack of explanation is not something we should sneer at. That is what conceptual and metaphysical primitivity is supposed to convey.

Original intentionality so conceived would then, of course, be part of the ultimate furniture of the universe. And so would minds. If they are not, then this is a bad mystery, and leaving it without an explanation is not an option for philosophers. This is how Jerry Fodor sees the issue:

> I suppose that sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of *spin, charm,* and *charge* will perhaps appear on the list. But

[110] The view that the semantic facts and properties cannot be told in a language that does not contain words that refer directly to such facts or properties is attributed to Brentano, and is often called the "Brentano thesis". A prominent present day defender of the irreducibility of semantics is Roderick Chisholm.

> *aboutness* surely won't; intentionality simply doesn't go that deep. It's hard to see, in the face of this consideration, how one can be a Realist about intentionality without also being, to some extent or other, a Reductionist. If the semantic and the intentional are real properties of things, it must be in virtue of their identity with (or maybe of their supervenience on?) properties that are themselves *neither* intentional *nor* semantic. If aboutness is real, it must be really something else.[111]

Philosophers who pursue this line of enquiry, of whom Fodor is perhaps the most prominent, usually end up saying that one thing being about another has to do with the former being caused by the latter. That is not very surprising. Physicalist ontology is not very generous. If there is really nothing but quarks, or maybe strings, and fields, whose catalogue of properties is exhausted by relative spatio-temporal positions and dispositions to enter into causal interactions under the laws of physics, there is not much of a choice if one contemplates to what he should try to reduce a relation between two things, which are not connected by relations of constituency. (And the physicalist's candidate for a mental representation, a neural pattern, I suppose, and the thing which it represents, a property "out there", do not seem to play any constitutive role for each other.)

Suppose the causal theorist of meaning establishes that some mental representations, I mean symbol types, are the representations of the property they stand for, *rather than of any other property*, because the occurrence of an instantiation of that property, under specifiable circumstances, reliably causes the occurrence of a token of the symbol type (in some or other faculty of the mind), while the occurrence of the instantiations of other properties doesn't (or only in a way that is "asymmetrically dependent" on the regular, veridical

---

[111] Fodor 1987, p. 97. An anti-realist about intentionality Fodor might have had in mind is Quine. Quine endorsed the thesis that intentionality is irreducible and, since he was a physicalist, drew the conclusion that then it cannot be real. Hence the term *semantic nihilism* that is often applied to his position. But even most dualists would say that Quine was too restrictive about the physical facts he considered eligible candidates for being the ones on which semantic facts could be reduced. He allowed only stimuli on nerve-endings and linguistic behavioural dispositions. This restrictive attitude he inherited form logical positivism and behaviourism, neither of which seem very plausible now to most philosophers.

cases of tokening the symbol). Or something like that.[112] If that much would be vindicated, part of the job would be done. The easy part. Maybe the causal theory can explain, and in perfectly naturalistic terms, why it is that a mental symbol is *about one thing rather than another*. But the tough question is why it is a symbol, *a mental representation* (at all, of anything), in the first place, *rather than a mere causal consequence*, which can have intentionality only in the derived sense. Why it is, in itself, about anything at all. In one sense, the level of mercury in the thermometer is about the temperature. But that sense is dependent on there being a mind, taking the level of mercury to be such a sign. Surely, under favourable circumstances, the level of mercury is appropriate for being used as such a sign. But imagine a totally mindless world in which the extension of some things varies with their temperature. Would their lengths be about the temperature? Or the other way around? Would anything be about anything else in that world? Now what the causal theory should be able to explain, and I don't think it has come anywhere near to it, is how it is that we have intentionality in a non-derived sense, in the sense in which the thermometer has not.

To this Fodor would of course reply that I vastly underestimate the significance of misrepresentations (when the instantiation of something other than the property which it represents causes the tokening of a symbol), which are asymmetrically dependent on veridical representations. It is them that make minds different from thermometers, they are responsible for the "jump" from the mere causal consequence to the symbol.

The level of mercury in the thermometer is a sign of the temperature—for someone who already has intentionality. But we are trying to account for how intentionality originally came to being, what the aboutness of the first thought on this planet, if you like, might have consisted in. How is that different form the aboutness of the thermometer? And here comes the answer: The thermometer cannot be mistaken about the temperature without ceasing to be a thermometer—without specific levels of mercury in it ceasing to be the causal consequence of specific temperatures, but we can take a horse for a cow without ceasing to be referrers—and without our Mentalese symbol "cow" ceasing to mean cow. Representations can go wrong, mere causal consequences cannot.

---

[112] That is the theory – with important refinements, of course – that Fodor advances, cf. ibid. pp.

How is it then that, if we take a horse for a cow, it doesn't bring it about, nor does it signal, that our Mentalese symbol "cow" means not cow but cow or horse? This is the "disjunction problem", and this is what the asymmetrical dependency of wrong tokenings on right tokenings solves.[113]

I see that the possibility of wrong tokenings does serve a function. Its function is to secure fixed meaning. And it does. The existence of wrong tokenings is a sufficient condition for the existence of fixed meaning. And asymmetrical dependence does solve the disjunction problem. But it is hard to believe that wrong tokenings are also necessary for fixed meaning. As it is hard to believe that the three conditions together are either necessary or sufficient for genuine underived intentionality.[114]

It is especially hard to believe that the capacity of getting things wrong could be the mark of true intentionality. There must be more to that. I would be very surprised to learn that it is something like a conceptual truth that someone who is warranted against taking horses for cows, God for example, cannot have genuine intentionality, because there is "no representation without misrepresentation". I am not assuming here that God exists. I am just assuming that it is not a conceptual truth that He—or some other perceptually perfect cogniser, Superman perhaps—can't exist. From Fodor's theory it would follow that God, if He exists, and if He is infallible, cannot really mean anything. And this is absurd. Or imagine a system consisting of a computer, a scanner, a software that recognises printed text on paper (a graphical image), when put in the scanner, as

---

[113] So the necessary and sufficient conditions of one physical entity "X" to be a representation (to stand for, to mean, to have the content, etc,) another physical entity X are these:

(1) X's reliably cause "X"'s.

(2) There are Y's, Y≠X, such that sometimes Y's cause "X"'s.

(3) For all Y's, Y≠X, if Y's cause "X"'s, then that is asymmetrically dependent on X's causing "X"'s.

This should be aboutness naturalised.

[114] One classic counterexample is Pavlov's dog. Substitute the salivation of the dog in place of "X". Food should be X, and the ringing of the bell Y. All three conditions are met. Does the salivation of the dog mean food (in the sense a thought means its content, with original intentionality, not just as a sign that needs to be used as a sign to mean anything by someone who already thinks)? Obviously not. To fend off such examples Fodor introduced a forth condition (Fodor, 1990), that the asymmetric dependency relation should be synchronic, not diachronic. It works against this example, but there are others.

a sequence of characters, a table on which all these stand, a room in which the table is, a building containing the room, and a tram, whose track draws on the street on which the building stands. Suppose that sometimes when a tram passes by the system takes an "o" for an "e", so the sequence of characters it produces misrepresents the printed text put in the scanner. It seems to be the vibration. Now such misrepresentations are asymmetrically[115] dependent on veridical representations. Fodor's conditions are all met. Does it mean that the system has original, underived intentionality? I find it hard to believe, but, for the sake of discussion, let's suppose that it does. Now imagine, that due to some reorganisation of public transportation in the city, from Friday next week no more trams traverse the street in front of the building ever. No more trams, no more "e"'s in place of "o"'s. No more misrepresentations, no more intentionality?

Or is it the potential to err that counts, even if it is not actualised any more in the lack of tram traffic? Then, what if a new edition of the software comes out, and equipped with that the system is not sensitive to the vibrations caused by (potential) trams any longer? It is a better version. Now is it sensible to assume that the system lost the property of original intentionality because some algorithms in the software had been improved? Again, it seems absurd.

I don't know of more promising attempts than Fodor's to reduce original, underived intentionality to the physical. And his attempt seems to be quite far off the mark.

But there is a philosopher who doesn't want to reduce intentionality directly to the physical, but who is happy to reduce it first only to the biological (and perhaps later, through chemistry, to the physical), and who claims, interestingly, that to be intentional we don't really need original intentionality.

Daniel Dennett in many of his works points to deterministic material mechanisms to which the "intentional stance" can be successfully applied.[116] If not thermometers, they can be patterns of cells in Martin Gardner's two-dimensional Life Game[117], or enzymes that perform a special duty in the controlling of sophisticated biochemical procedures, or chess-computers, or robots that were programmed to adapt to their environment (avoid harm, find

---

[115] And the asymmetrical dependence is synchronic.

[116] It doesn't change anything important if some parts of these mechanisms are genuinely indeterministic.

[117] Gardner 1970.

resources, co-operate or compete with fellow robots) and survive until 2401, because their creators who wait inside them hibernated wanted to see what the 25th century would be like, and so on.[118] The successful application of the intentional stance means basically that although these mechanisms all have complete descriptions in terms of their physical constituents and the rules, laws or algorithms that, relative to the environment, prescribe what should happen to them, which are *totally blind to either meaning or purpose*, it is possible to describe their history as if they represented and understood the environment and acted on purpose relative to that understanding. The application of the intentional stance may resemble actual successful scientific theories by accounting for existing data on the behaviour of the mechanism in question and by making reliable predictions. The fact that there is a deeper non-intentional description of the system (non-intentional in the sense that on that description nothing in the system is about anything in the environment, and also the system doesn't intend to do anything) does not invalidate the intentional account. It is like when we started to handle quantum mechanical multi-body problems with perturbative methods relatively successfully, yet this new knowledge did not invalidate the theories of phenomenological chemistry. They just received a deeper explanation. Phenomenological chemistry didn't go as the phlogiston did, and the intentionality of these systems is not like the phlogiston either. In many cases using the underlying physical description rather than the intentional description for predicting the system's behaviour is inadvisable. For it may be hopelessly demanding computationally and the gain in terms of the reliability of the prediction may be vanishing relative to the excessive effort.

I think (with many naturalists, Fodor included) that the "intentionality" of these systems is intentionality only in the metaphorical sense, or in the derived sense. Paradigmatically, the intentionality of the enzyme in the metaphorical sense[119] (but of

---

[118] The example of life games is extensively used in Dennett 2003, that of the chess automat in Dennett 1990, the enzyme and the robot in Dennett 1988.

[119] Look at the passage about the "intentionality" of macromolecules Dennett quotes from Robertson: "... A much more demanding *task* for these enzymes is to *discriminate* between similar amino acids. ... However, the observed *error* frequency in vivo is only 1 in 3000, indicating that there must be subsequent *editing* steps to enhance fidelity. In fact the synthetase *corrects* its own *errors*. ... How does the synthetase *avoid* hydrolyzing isoleucine-AMP, the *desired* intermediate?" (pp. 664-5; Rosenberg's original emphases.) These must be clear cases of "as if" intentionality.

course that is derived, too, derived from the intentionality of the one who coins the metaphor), that of the robot in the derived sense.[120] And I think (with Fodor) that the naturalist should want more. He should aim at explaining original intentionality.

Dennett doesn't debate that these are only cases of derived intentionality. But he thinks it is a mistake to want more. He invites us to believe that we ourselves are like the cell pattern, the enzyme, the chess automat, or the robot. He thinks that this is the best kind of intentionality one can get, for the very simple reason that this is the only kind that exists.

However, describing the "intentionality" of material mechanisms, ranging from the enzyme to the robot, as a property *attributed to them* from an explanatory perspective, *and* claiming at the same time that this is the only kind of intentionality seems to be a contradiction. The intentional explanatory perspective vanishes if there is no one whose perspective it would be. Or is there someone? Is the idea that some mechanisms to which the intentional stance is applicable can themselves apply the intentional stance, to other things and themselves? It cannot be. For surely it cannot be the explanation for how the intentional stance came to the world (the explanation for the first thought with content). For there being an intentional stance, there must already be someone who thinks, i.e., has intentionality. So his intentionality cannot just be the intentionality attributed to him applying the intentional stance. Whose intentional stance would that be? Depending on the answer (himself or someone else), this explanation of intentionality is either circular, or faces an infinite regress.

But Dennett seems to have a third answer. It is the intentionality of Mother Nature from which our intentionality is derived:

> As a late and specialised product, a triumph of Mother Nature's high tech, our intentionality is highly derived, and in just the same way that the intentionality of our robots (and even our maps and books) is derived. A shopping list

[120] It is not the robot that represents the environment, it is its creator who uses some states of the robot to represent some states of the environment, the aboutness of any state of the robot is derivative of the original intentionality of the thoughts of the creator, or maybe of the intentionality of our thoughts, who are now accounting for the robot's behaviour under the intentional stance. And the same applies, of course, to the purposes of the robot, as well.

> in the head has no more intrinsic intentionality than a shopping list on a piece of paper. What the items mean (if anything) is fixed by the role they play in the larger scheme of purposes. We may call our own intentionality real, but we must recognise that it is derived from the intentionality of natural selection, which is just as real—but just less easily discerned because of the vast difference in time scale and size.[121]

But here I am seriously puzzled. I thought that one of the beauties of the evolutionary theory for naturalism was that it promised to explain our design features without positing a designer with intentions (who intended it that we be they way we are)[122]. And now I hear that our intentionality is derived and is derived from the intentionality of Mother Nature (or national selection, or our genes, see the previous footnote), and Her intentionality is real (just difficult to discern, because we are small and She is big). Do we have an intentional designer, or don't, after all, according to Dennett? Or is it that I shouldn't pose the question like this, because he is talking metaphorically when he talks about the intentionality of Mother Nature, or of genes, and the like. Probably so. What can be more stupid than a gene? A gene surely doesn't mean anything. But then

---

[121] Dennett 1987, p. 318. At another place (1990b, p. 59.) he says it is the intentionality of our genes from which our intentionality derives: "We now have an answer to the question of where we got our intentionality. We are artefacts, in effect, designed over the aeons as survival machines for genes that cannot act swiftly and informedly in their interest." (So this is how we are analogous with the robot that was designed to take its designer to the 25th century. Dennett continues:) "So our intentionality is derived from the intentionality of our "selfish" genes. They are the Unmeant Meaners, and not us..." (The "Unmeant Meaner" is the one endowed with original intentionality that is not derived from the intentionality of anyone else who applies the intentional stance to him (or it?). The regress of explaining intentionality with reference to the intentional stance stops at the Unmeant Meaners, as the exploration of the causal history of whatever is taking place was once thought to stop at the Unmoved Mover, in order to avoid an infinite regress of causes.) "...and in so far as some theorist can interpret an event or structure in us as being about something or other...it is only because of the informative role that such signaling plays within the artifact, and the way it contributes to its self-preservation."

[122] I mean, I thought that one of the values of evolutionary theory for physicalism is that it explains how Paleyanism can be wrong – how is it that there can be functional biological structures, more complex than a watch, from which we don't need to infer the existence of a designer and creator, more intelligent and powerful than a watchmaker. Or, to use Richard Dawkins's metaphor, how is it that the "watchmaker" can be "blind", that is, devoid of intentionality.

why is he saying that this "intentionality" is real? And if it is not real, just metaphorical, then our derived intentionality is derived from a metaphorical intentionality. Whose metaphor is it? I thought it was us who described the workings of Mother Nature from the "intentional stance" and endowed her with intentionality, of course, metaphorically. Then isn't the derivation of our derived intentionality going to be circular, after all?

Maybe I completely lost track of when Dennett speaks metaphorically and when he really means what he says. Or, and this is what I suspect, what he says is incoherent. Or, this is another possibility, he is an antirealist about meaning, after all. He is a realist only about syntactical machines that produce responses to environmental input. Some syntactical machines produce responses that serve their survival well, others don't. The latter die out. Does that mean that the former understood more of the environment? In one sense, yes. In the metaphorical sense. Is that all that there is to say about understanding? I don't think so. Being a fit syntactical machine doesn't endow one's syntax with semantics. And neither does mere causal relatedness to the environment (or so I argued, reviewing Fodor).

So every single thought seems to be a refutation of the thesis that there is only metaphorical intentionality attributed to us applying the intentional stance.

Ruth Millikan suggested, contrary to Fodor, and to all naturalists who hope to give an account of original intentionality in causal terms, that the supposition that such intentionality exists is incompatible with naturalism.[123]

I think she is right. We just draw the opposite conclusions.

That much about item (i) of my list of elements of justifiedness. Of course, if my conclusion is correct, then it resounds on the whole list. Because from item number (ii) every item on the list has to do with normativity and truth, which is intertwined with the issue of aboutness.

(ii) is about the capacity to discern cases when something is represented as it is, from cases when it isn't. Very trivially, if aboutness is not real, then no thought is really a representation of anything, and then the whole issue of the truthfulness of the

[123] Millikan 1984.

representation, i.e., whether it corresponds to the bit of reality that it is supposed to be about, simply does not arise.[124]

If aboutness is real, and a strictly physicalist ontology cannot accommodate it, that is enough reason to dismiss physicalism. Of course, if a world cannot accommodate aboutness, it cannot accommodate truth either.

Quine, who thought that aboutness is not real, famously claimed that naturalised epistemology can do without the idea of truth. His idea was to get rid of normativity altogether. This is how he saw epistemology after it's been achieved:

> Epistemology still goes on, though in a new setting and a clarified status. Epistemology, or something like it, simply falls into place as a chapter of psychology and hence of natural science. It studies a natural phenomenon, viz., a physical human subject.[125]

This chapter of psychology, this study of human subjects, is the study of how they come to believe things. Some ways of coming to believe things are coming to have knowledge. The task of epistemology is to describe these ways. But if we fend off normativity, then these mechanisms of acquiring beliefs cannot be "reliable" in any sense that would require that they reliably produce beliefs that are true, or correct, or adequate. But then how to distinguish between reliable and unreliable ways of belief acquisition? Or are we just to describe different ways of how beliefs are produced in us, without holding any of these ways better, in any sense, than any other? But then how this *describing* of belief acquisition would be different from any random talk of belief acquisition? What makes it a description of anything, if not matching up with some reality, in some sense of adequacy or truthfulness?[126]

---

[124] I am assuming the correspondence theory of truth. I am a realist. This is an assumption for which I am not going to argue within the limits of the present work.

[125] Quine 1969, p. 101.

[126] This point is of course not original with me (cf. Willard, 2000). Putnam too found naturalising epistemology by way of denying normativity entirely hopeless on similar grounds: "Why should we expend our mental energy in convincing ourselves that we aren't thinkers, that our thoughts aren't really about anything, noumenal or phenomenal, that there is no sense in which any thought is right or wrong (including the thought that no thought is right or wrong) beyond being the verdict of the moment, and so on? This is a self-refuting enterprise if there ever was one!" (Putnam 1983, p. 246, quoted by Willard, 2000.)

Maybe Quine was wrong to think that the issue of naturalising epistemology and the issue of normativity are so entangled. But either way, naturalism seems to be facing serious problems here. If Quine was right, then naturalists have to do without the distinction between a thought that is an adequate representation of a fact, and another that isn't, and this seems absurd. If Quine was wrong, however, and naturalist accounts of epistemic states can involve reference to normative notions like truth or correctness, then it is hard to see how that would fit into their narrow ontology, which doesn't seem to provide for the relation of one thing representing the other, let alone representing it correctly.

Needless to say, it affects every item on the list from (ii) to (x). If there is no normativity, then there is no difference between those who infer correctly and those who don't. Because then there are no logical norms. So it is hard to make sense of conceptualisations that would serve the purpose of generalisations, because they are dependent on logical relations. So there is no inductive thought. There is no deductive thought either, for obvious reasons. And, of course, there is no noetic unity, whose cement seems to be logic, too.

One might protest that these points concerning aboutness and normativity are not exactly Epicurean points, and they divert the discussion from its original topic. To some extent, this is a correct observation, since these points have nothing to do with determinism, or with the alleged passivity enforced on cognisers by determinism. It is a problem for physicalist ontology.

However, in chapter 2 I argued that there is little reason to believe in determinism if not for the reasons that I presented there. Those reasons were the thesis of physical determinism, in combination with the thesis of the causal closure of physics, that together yielded universal determinism encompassing human thought, or at least the part of human thought that plays a role in action production. We have seen that the causal closure thesis, if true, yields a powerful argument for physicalism in the form of a psychophysical identity (or supervenience) thesis. Determinism and physicalism are, of course, logically independent, but I doubt that anyone would be able to point at any evidence that would suggest that we operate deterministically, assuming that our mental life is genuinely non-physical. There are, of course, some findings of behavioural psychology about types of agents, types of mental conditions and types of circumstances that

make types of actions probable. But these are way off the mark, if establishing determinism as a general thesis about human behaviour is the issue.

So naturalism, if meant as involving the assumption of a narrowly physicalistic ontology, and determinism are not independent of each other, after all. For it is the naturalistic metaphysical background theory that renders determinism plausible. Determinism, not logically but argumentatively, is dependent on this theory, at least at the present stage of our knowledge about ourselves and the world. So if this background theory is found incompatible with subjects having certain mental states, then the existence of those mental states decrease the plausibility of determinism.

So this is an Epicurean concern, even if in a broader sense, after all.

And it is also important to note that philosophical ideas usually do not stand on their own. They attack in strong alliances, mutually covering each other's back. As you already might have noticed, I have a tendency to think of compatibilism about freedom and determinism as a member of such an alliance. In my view, it is part of a broader naturalist agenda. I think compatibilism, at bottom, is the project of naturalising freedom, alongside other projects, such as naturalising morality, naturalising epistemology, naturalising intentionality or mental content in general, and, one day, consciousness altogether, and everything that once seemed to be irreducibly mental or personal, and, as such, non-physical. The above considerations, if correct, weaken the alliance, the broader philosophical agenda, of which compatibilism is a part. So they are relevant to our topic also in this capacity.

The issue of passivity

But of course the argument I gave for the claim that intentionality cannot be naturalised was inconclusive. My argument consisted of two components essentially. One: pointing out that intentionality is puzzling and the puzzlement disappears easily only assuming the irreducibility of intentionality and dualism, and two: discussing two particular attempts to naturalise intentionality, Fodor's and Dennett's, and recording that both are unsatisfactory, plus that there is not much

more on offer at the market.[127] But of course I did not and I cannot exclude the possibility that one day someone comes up with a real solution to the riddle of squaring intentionality with physicalism.[128] So we have to consider what if item (i) on the list of the elements of justifiedness is cleared by the determinist-physicalist theorist. What if

---

[127] But see Dretske 1981 and Millikan 1984.

[128] There is an interesting article by Stephen Stich and Stephen Laurence (1994) in which they argue that even if it is true that intentionality cannot be "naturalised" in any usual sense of the word, it doesn't mean that in a naturalistic world nothing can instantiate intentional properties. They consider different variations of what "naturalisation" could consist in, i.e., giving necessary and sufficient conditions of something being about something else in a non-intentional, physical vocabulary, or specifying, in non-intentional terms, an underlying clearly physical property which is always instantiated when aboutness can be predicated of something (which would mean that intentional predicates are natural kind terms), or specifying non-intentional, physical properties on which intentional properties supervene. They argue that the conclusion that intentionality has no place in a physicalistic world doesn't follow from the failure of either of the first two of these three ways of naturalising. Then they consider three possible candidates that might be proposed as the supervenience base of intentional properties: a) the current, internal physical properties of the organism in question, b) the physical properties of the organism (dropping the requirements that they should be current and internal), and c) the class of all non-intentional, physical properties (dropping the requirement that they should be instantiated by the organism). They claim that if someone would prove that intentional properties do not supervene on either of the first two bases, the conclusion that one has to choose between meaning irrealism or giving up physicalism would not follow. Finally they claim that it is impossible that intentional properties do not supervene on the third proposed base, because if x and y share all their intrinsic and relational physical properties, their spatiotemporal location, and their history, then they are identical, therefore they must share their intentional properties, as well. I do not want to discuss their argument in detail here. I think they are probably right up to the last point, but their final argument is wrong. Nevertheless there is an important lesson to be drawn from the article, and this is that the failure of naturalising efforts, even the proven failure of all possible naturalising efforts of some particular kinds (except, of course, for the last kind – supervenience on the widest possible physical, non-intentional base) doesn't show that meaning is either unreal or genuinely non-physical. Admitting this, if their last argument fails, as I think it does, then it would be a misrepresentation of the dialectical situation to say that the remainder of the argument shows that the critics of Fodor or Dennett or any other particular efforts to naturalise intentionality *failed to show* that intentionality has no place in a physicalistic world. I think without the final argument their article shows that unless the dualist proves that supervenience on the widest possible physical, non-intentional base is impossible, it should be considered possible despite of the failure of the discussed naturalising attempts. But it doesn't mean that it should also be considered as plausible. There is something initially puzzling about intentionality for the physicalist, as Putnam and Fodor has pointed out. It is the physicalist theorist who *has to show something*. Unless something is positively said about how intentional properties would supervene on non-intentional ones, the puzzlement remains.

a deterministic material mechanism can have intentional states? Can he also perform the other tricks on the list?

And now come the problems that are linked directly with determinism.

There are points of the justificatory process at which practical deliberation is embedded in it.

One obvious such point is when we decide whether the empirical evidence for an inductive generalization is strong enough for us to embrace the generalization (item (v) on the list). Shouldn't we keep on gathering evidence? Shouldn't we try to think of new places to look for possible counterexamples? These questions may fade away quite automatically as we keep on checking the world for possible counterexamples, but in many cases, in science in particular, a conscious decision terminates this stage and leads to the next when we do not hold ourselves apart from the general empirical thesis any longer.

A little less obvious, yet obvious enough, is that we might deliberate about how to form a general concept. We might come up with a conceptualization with which a general hypothesis involving the concept that seems appealing to us gets falsified by counterexamples. In such cases we might review the logical links our concept bears to others, and coin a slightly different conceptualization with largely the same content, with which the generalization works. It seems that in this process we deliberate about how exactly we should define a concept, in terms of the logical links it should bear to others, so that it might serve best our interest of mapping a part of the world.[129] (This is item (iv) on the list.)

One might think that maybe there is some deliberation involved in inductive reasoning, but if our reasoning is clearly deductive then deliberation plays no role in it. Very truly, if we hold true some premises and see the entailment, then there is not much to deliberate about whether or not we should hold true the conclusion. There

---

[129] A classic illustration of this point is Hilary Putnam's account on the evolution of the concept of kinetic energy in his "The Analytic and the Synthetic" (Essay 2 of *Mind, Language and Reality*, 1975). Putnam's point is that some conceptual truths, that *prima facie* seem to hold by stipulation, like the one that by "kinetic energy" we should mean the half of mass time speed squared, are revisable under the weight of empirical data, as it happened in the relativistic modification of the concept of kinetic energy by Einstein, as it turned out that the modified concept (momentum squared divided by two times the mass) serves better out need to grasp a feature of reality.

would be nothing wrong about being logical machines that produce true conclusions from true premises automatically.

But we are not such machines. Most of us draw deductive inferences routinely, automatically, sometimes even half-consciously. However, most of us, I guess, have enough experience to know that sometimes deductive inferences should be checked twice. Sometimes it seems necessary to make a deductive inference fully conscious and transparent to scan it for possible mistakes. In such cases we deliberate first about whether the inference should be checked (whether there is reason to suspect it might have gone wrong, or whether it is so important that it needs to be double-checked anyway, even if there is no such doubt), and then, when we decided to check it (along ways described in point (viii) of the list), then we deliberate about whether we find it that the inference has been cleared from any possible doubt, and the checking process can be terminated. (In theoretical contexts sometimes we go to the extremes to get clear of doubts—the case of the consequence argument in chapter 2, a deductive inference which *prima facie* seems straightforwardly true, is a good example.)

Having said all this, it should be clear that the final judgement about a thesis (of a level of abstraction and generality comparable to that of the thesis of determinism), based on these intermediary judgements, also involves an element of practical deliberation. (This is item (ix) on the list). We do more than just remembering the previous judgements we made along the way. The more components of the justificatory process had that involved an element of practical decision, the more reason we have for caution committing ourselves to the truth of the proposition in question. We might decide to perform a final round of checking. We may do this by anticipating objections to the proposition. We also might want to check how it fits with the rest of what we believe (see point (x) on the list). There seems to be a practical decision about when this holding ourselves apart from the proposition should be terminated in a commitment.

These elements of practical deliberation and decision intuitively seem to have a different meaning depending on whether the decisions are made freely in the libertarian or only in the causal (compatibilist) sense. If we are only compatibilist free, then all our assents to propositions are such that it was never really possible that we would conceptualize differently, decide to go and check other areas of reality for the truth of our empirical premises, decide that the affirmative

evidence we had for them was not yet robust enough, that we should put even more scrutiny on our deductive inferences, or hold back from the final commitment a little longer. Then we were *caused* to regard our doubts to have been put at rest exactly when we did so. And there seems to be something worrying about having been caused to terminate deliberation about whether we should hold something true or not. The worry is that then this decision is not ours after all. The propositions we believe are not those *we freely decided* to take to be beyond reasonable doubt, so to speak, but those that were *thrashed into us* by those causes.

In some contexts when the freedom of the will is concerned, and reasons that play a role in will-formation are modelled as desire-belief complexes, we tend to say that it is our desires we want to be free about, not our beliefs, since it is all right if our beliefs just track the truth, we shouldn't desire the freedom of having false beliefs. I used this reasoning, too, in support of a useful simplification of our picture of our volitional hierarchy in a footnote in chapter 3. Now it seems that it was an oversimplification. It seems that we should desire some freedom in respect of what we believe, even if we desire to believe only what is true. The function of this freedom is to have the power to hold back from committing ourselves to the truth of some propositions (which, without this freedom, we would perhaps be caused to believe) until all doubts about them is cleared.

In respect of many things we believe there don't seem to be natural thresholds when to terminate the practical deliberations that are embedded in the justificatory process, at which we could point and say that it is all right if we are caused (programmed) to stop deliberating exactly at those thresholds. On the contrary, many of us seem to have very different thresholds for the same question (for example, David Lewis found it necessary to consider the possibility that there are local miracles before assenting premise 2 of the consequence argument, while, I suppose, some of us with comparable general intellectual cautiousness would have found it unnecessary), and each one of us seems to have very different thresholds for different questions. So maybe we should desire to be free to decide if the justification we have for particular propositions that suggest themselves to believe satisfy *our intellectual conscience* that has its own standards, maybe different standards for different contexts. If determinism is true, then this freedom can be given only a quite vacuous sense.

This is one issue related with determinism and passivity.

The Lewis – Anscombe debate

But there is an argument that is not dependent on there being practical elements in theoretical reasoning. It is not a completely different argument, though. It is a more general argument, of which the problem with practical deliberation embedded in justification is a subcase. It concerns reason in general, both practical and theoretical. It is the general version of the Epicurean Argument, and it had many lives. It was reborn with Kant, interestingly with a British prime minister who had an interest in philosophically defending the Christian religion, Arthur James Balfour, in the beginning of the last century[130], and entered the modern philosophical scene with a famous debate between C. S. Lewis and Elizabeth Anscombe at the Oxford Socratic Club in 1948.[131]

The intuition on which the argument is built is very simple. As Lewis stated it in the 1948 edition of *Miracles*, the idea is that *"no thought is valid if it can be fully explained as the result of irrational causes"*[132]. According to Lewis, the combination of determinism and materialism is exactly the hypothesis that all thought is fully explained as the result of irrational causes. So if both determinism and materialism are true, then no thought is valid. Including the thought that determinism and materialism are true. So the conjunction of materialism and determinism is either false, or cannot be thought validly ("validly" meaning "reasonably", we may suppose).

Lewis's own wording is a little different:

> It would be impossible to accept naturalism itself if we really and consistently believed naturalism. For naturalism is a system of thought. But for naturalism all thoughts are

---

130 Balfour 1989.

131 As far as the historical matters of the Lewis-Anscombe controversy is concerned I rely largely on an on-line paper by Steven Lovell, which he published on the Internet alongside his PhD dissertation entitled "Philosophical Themes from C. S. Lewis", which he defended in the Department of Philosophy at the University of Sheffield in 2003. The paper is entitled "C. S. Lewis's Case against Naturalism" (Lovell 2003). Unfortunately, there is no indication that Dr. Lovell has published his paper in a more conventional format. In some places my discussion of the Lewis-Anscombe debate follows his. I will indicate whenever a point I owe to Dr. Lovell.

132 Lewis 1948, quoted by Lovell.

> mere events with causes. It is, to me at any rate, impossible to regard the thoughts which make up naturalism that way and, at the same time, to regard them a real insight into external reality. … Every particular thought … is always and by all men discounted the moment they believe that it can be explained, without remainder, as the result of irrational causes. Whenever you know that what the other man is saying is wholly due to his complexes or to a bit of bone pressing on his brain, you cease to attach any importance to it. But if naturalism were true, then all thoughts whatever would be wholly the result of irrational causes. Therefore, all thought would be equally worthless. Therefore, naturalism is worthless.[133]

Lewis makes no explicit reference to determinism. But from the text it is clear that he takes determinism to be part of the doctrine he calls "naturalism". What he had in mind, under the label of "naturalism", is what we earlier called "mechanistic materialism", which is the conjunction of three theses: psychophysical identity or supervenience (ontological physicalism), determinism, and the thesis that physical causes are mechanistic in the sense that they are blind to any purpose or meaning. That the latter two theses are included in what he calls "naturalism" accounts for the fact that he takes naturalism to include the thought that all thoughts are caused, and that the causes are "irrational". The conclusion, "naturalism is worthless", I think we can take to mean that naturalism is either false or unreasonable.

Elizabeth Anscombe made two powerful objections to this argument.[134]

Firstly, she pointed out that Lewis has conflated the non-rational with the irrational, and claimed that the argument trades on this conflation. The naturalist (I am using the word now in Lewis's sense) is committed to the view that the causes of thoughts are just like the causes of other occurrences: physical, and as such, devoid of meaning or purpose, bearing no intrinsic rationality. It doesn't mean, however, Anscombe thinks, that they should be lumped together with irrational (intrinsically rationality-diminishing) causes, like complexes or bits of bones pressing on bits of brain. Irrationality is not the same as non-

---

[133] Lewis 2002, p. 170 (the text is originally from 1946), quoted by Lovell.
[134] Anscombe 1981.

rationality. The physical causes of mental effects the naturalist is assuming are not in themselves irrational, rather, they are occurrences to which the rational/irrational divide does not apply. Being non-rational means exactly this.

The objection is fair; Lewis was indeed a little careless about choosing his words. Non-rationality is different from irrationality. But I am not sure if the argument trades on the conflation of the two. What Lewis had in mind could have been stated like this: For the reasonability of a thought it is a *necessary* condition that the process whose product it is have the quality that the rationality/irrationality distinction is applicable to it. Being the product of such a process is of course not sufficient for the thought's being reasonable. (The process can be irrational, and then the thought is not reasonable). But the point is that it is necessary. If a thought is a product of a non-rational process, then it cannot be reasonable. The problem with the alleged physical causes of our thoughts is exactly that they are blind to the rational/irrational distinction. That this was how Lewis saw the matter is reflected in the fact that, although the general perception was that he suffered an overwhelming defeat from Anscombe at the meeting of the Socratic Club on 2 February 1948, in the next, revised edition of *Miracles* he left the argument practically unchanged, just replaced references to "irrationality" with references to "non-rationality".

But Anscombe's other objection attacks Lewis's argument at a point, where it seems really vulnerable, or at least in need of substantial support from further argument. Anscombe points out that Lewis relies on a hidden premise, which he takes for granted, although it is reasonably debatable. The hidden premise is that a rational and a non-rational explanation cannot simultaneously be true of the same event, if both are meant to explain the event fully. Unless we assume the truth of this extra premise, the argument doesn't go through, because, if this is false, then the fact that a thought was the effect of non-rational causes may sit well with its being perfectly reasonable. So the argument turns on the truth of this extra premise.[135]

---

[135] Anscombe made also a third objection, which, however, seems entirely unconvincing. She said that Lewis's claim that if naturalism were true then no thought would be reasonable, for no reasoning would be valid, is meaningless, because validity, or the lack thereof, cannot be claimed of human reasoning in general. This is so, according to Anscombe, because we acquire the concepts of validity and invalidity from encounters with particular reasonings that are valid and others that are invalid. If naturalism really

Anscombe suggests that we shouldn't accept this extra premise because explanations compete only if they belong to the same kind, and causal and rational explanations belong to different kinds.[136] Steven Lovell quotes a passage from Anthony Flew, one of Anscombe's allies in the debate, making essentially the same point.

> Lewis and others who produce similar arguments are snared by the chronic ambiguities of words like "cause," "reason," "because." If asked "What is the reason why you think this is true?" I may reasonably answer either "It was thrashed into me at school," or "It follows from such and such true premises." Both these answers simultaneously may be sound, for they are answers to what are really quite different questions. I shall call the senses of "reason," "cause," etc., which ask for the first type of answer the *historical* senses…, and shall call the senses which ask for

---

entailed that no thought is valid, then it would also entail that we cannot even have the concept of validity. Up to this point Anscombe seems right. But it is hard to see why Lewis couldn't embrace this idea and claim that the mere fact that we have the concept of validity refutes naturalism. Anscombe says that this position is not available to Lewis, because the talk of ourselves as completely lacking the concept of validity is incoherent, because the mere thought refutes its content. But it doesn't seem to work. Undoubtedly, Lewis can meaningfully talk about dogs lacking the concept of validity altogether. There is no contradiction in it as long as Lewis thinks that he, or humans in general, has or have that concept. Lewis could not coherently claim of himself, or of humans in general, that he lacks, or we all lack, the concept of validity. But he can meaningfully claim of himself (or of humans in general) that he has (we have) the concept of validity, or, which is the same, that it is not true that he (we) lack the concept of validity. Similarly he can coherently conceive and talk of counterfactual situations in which he (we) lack the concept. What he cannot coherently think or claim is that the actual situation is such. He can think or talk coherently of situations in which naturalism is true, supposedly bearing the consequence that then we cannot have the concept of validity, as long as those situations are thought to be counterfactual. What would be incoherent is to think or talk of these situations as real or actual. That naturalism cannot be claimed coherently because it bears the consequence that we don't have the concept of validity, and one cannot coherently claim that we actually lack it, since such a claim would refute itself—this could be one version of Lewis's argument. Surely, Anscombe cannot refute Lewis this way. If this objection of Anscombe's were right, then there would be no indirect proofs. An indirect proof for *p* asks us to consider what if non-*p*, and proves that something contradictory to already accepted premises follow from non-*p*. One could object against such a proof that at some point of the proof we predicate an incoherent set of predicates of a situation, namely the situation when non-*p* is true, and an incoherent set of predicates cannot be predicated of anything, so the proof cannot be right. But surely, this is not a good objection. There are indirect proofs. (Lovell comes to the conclusion that this objection of Anscombe's is unsound a different way..)

[136] Anscombe 1981, p. 228.

> the second type of answer the logical *senses*…. If the reason (historical) why I think my mental processes are determined by neurone changes is itself something to do with neurone changes, this has no necessary bearing on the questions whether there are, or whether I have, any logical reasons, any good arguments, for thinking this thought about the causation of my mental processes.[137]

But it is too simple. We should distinguish between different senses of "because". One distinction to be drawn is between the "because" of rational explanations that are mere *post facto* rationalizations, and the "because" of those that are more than that. A mere *post facto* rationalization does not compete with a causal explanation.

But what makes a rational explanation more than just a *post facto* rationalization?

I think what distinguishes between the two is that a mere *post facto* rationalization is not a true account of why the cognizer came to think the thought that is being explained, while a rationalization that is "more than that" is. Consider that a rationalization invoking a reason that the cognizer really considered, and really favours the thought that is to be explained, can be a mere *post facto* rationalization, if it is not true that it was the reason for which the cognizer adopted the thought. There is something "historical" about the criteria that distinguish between true rational explanations and mere rationalizations.

But one may try to insist that even though both rational and causal explanations are "historical", they are "historical" in different ways that are not competing with each other. To show that this is possible the adherents of Flew's position should demonstrate that the property of a reason that it was in fact the reason for which the cognizer adopted the thought that is being explained can be accounted for in non-causal terms, and this account does not exclude a causal explanation for the same thought that doesn't cite the reason among the causes.[138]

---

[137] Flew, "The Third Maxim", *The Rationalist Annual,* 1955, p. 65.

[138] At this point what comes to everybody's mind first is Donald Davidson's classic argument, in the context of explaining actions, rather than thoughts, to the effect that the account for a reason's being the one for which the agent actually acted cannot be non-causal (1963). Davidson's considerations can be directly transplanted from the practical context to the theoretical one. Suppose I believe in God. Suppose I can give

I think it is impossible. But it is not necessary to prove that it is impossible to see that the Lewisian conclusion cannot be escaped this way.

Suppose for the moment that this is possible, and the situation is as Flew suggested: a rational explanation invoking the reason that was actually effective and a logically and metaphysically independent causal explanation can happily coexist. What would it mean for a reason to be effective, if there was an independent causal story to be told about how the thought to be explained had come about? In such a case the reason, that is, the *rational ground* for the thought, is not necessary for the thought to be thought with assent. The cognizer could have thought it with assent in the lack of the rational ground just as well as in the presence of it, since the causal mechanism to which the causal explanation refers, if it is really independent of the rational explanation, could be there unchanged even in the lack of the rational explanation, and it could bring it about that the thought is thought with assent on its own. But then, in the presence of the appropriate non-rational causes, the cognizer couldn't have refrained from thinking it with assent even if there was no rational ground for it at all.

It doesn't necessarily prevent the cognizer from seeing that there is no rational ground for what he is thinking. (If determinism holds,

---

strong arguments in favour of the existence of God and against a godless, say, materialistic worldview. Suppose I do entertain all those thoughts that may serve as the rational ground for believing in God, and I see that they are a rational ground for being a believer. But it is not true that that is why I am a believer. I fear that my existence is meaningless and death destroys it completely and forever. This existential anxiety makes me believe in God, because this is the only thing in which I can find refuge from this anxiety. I am clever enough to provide a rationalization. The rationalization may be sound, yet a false account of why I am a believer. Davidson thought that, in practical contexts, the difference between the reason on which an agent acted and another reason which the agent considered, can be used to rationalize what he did, but which is not the one on which he actually acted can be accounted for only in causal terms: the reason on which he acted caused the action, the other reason didn't cause it. The same consideration would suggest here, in the theoretical context, that the difference between a reason that caused me to think that God exist, and a reason that merely rationalizes my faith can be distinguished only causally. I think, however, that in both contexts, practical and theoretical, Davidson's conclusion is too strong. It is not true that the effective reason must be a cause, but it is true that a reason's being effective excludes an independent causal explanation. The effectiveness of a reason can exclude an independent causal explanation even if it is not a cause, if a reason's being effective requires a judgement or decision to assent to, or act in accordance to, what is suggested by the reason, which is not itself caused by anything but which interferes with the causal order.

whether he sees that depends on whether there are causes that make him see it. There may be such causes. But of course, there may also be causes that make him think that there is such a rational ground although there isn't.) But it makes this suggestion even more inadmissible. For then it would be possible to see clearly that one has no rational ground for thinking something, and yet being fully committed to the truth of the thought, if one is caused to be in that state of mind. And this, I think, completely undermines rationality.

On this account we think what we are caused to think. And if causes are logically and metaphysically independent of reasons, what Flew's claim that rational and causal explanations are answers to two equivocated but entirely different questions seems to amount to, then it is a matter of pure luck if there is also a rational ground for it. Flew's thesis of course doesn't exclude that there be also valid reasons for what we are caused to think. But our thinking what we think is not dependent on it. If a thought is rational, that is a lucky epiphenomenon. Reason doesn't interfere with the causal order, and so we have no power to resist irrational thoughts if we are caused to think them.[139]

Flew was aware of this problem, and he proposed a solution: evolution. Although it is, in principle, possible to be hooked up causally so that one thinks with assent thoughts for which one has no rational ground, we, humans, are hooked up so that a significant portion of what we are caused to think is also rationally grounded, and this is not a mere coincidence. This is a feature for which our species was selected.

More precisely, this is not exactly what Flew said. He said that evolution can explain why it is not a coincidence that a significant portion of what we are caused to think is *true*.[140] But if only this was true, that would not save rationality, for the rationality and the truthfulness of our thoughts are not the same.

The difference is exactly the same as the difference between knowledge on accounts that don't require justification, and knowledge on accounts that do. The difference is rational justification, on the ground of which one commits himself to the

---

[139] Lovell develops essentially the same point drawing on a passage from Ernest Gellner debating with Flew in *The Rationalist Annual* (Gellner 1957, "Determinism and Validity", p. 74.).

[140] Flew offered this argument in response to Gellner in the next issue of *The Rationalist Annual* (Flew 1958, "Determinism and Validity Again", pp. 46-7.) Quoted by Lovell.

truth of some propositions in an intellectually responsible way. Evolution may select us for representing our environment adequately, reliable causal belief-producing mechanisms may have a survival value over non-reliable ones, which may account for why the former will be the prevailing traits, and why the latter die out. But if the causal and the rational explanation for the same belief are non-related the way Flew claims they are, then the presence or absence of a rational justification does not interfere with the causal mechanism, which is then exclusively responsible for our thoughts, and then the presence or lack of a rational justification is completely invisible to evolutionary selection. Being armed with reliable belief-producing causal mechanisms, plus having accompanying rational justifications for a good deal of what we are thinking has no extra survival value over just having the reliable mechanisms without accompanying rational justification. Evolution may explain why it is that a significant portion of what we think happens to be true even though we are caused to believe them by non-rational causes, but it cannot explain why we are rationally justified in a good part of what we truly believe, if rationality and causality are related as presented by Flew.[141]

---

[141] Alvin Plantinga has an interesting argument that is purported to show that even reliable belief-producing mechanisms, not to mention accompanying rational justification, cannot be the products of natural selection (of evolution in a naturalistic world in which evolution is guided solely by causes that know no meaning or purpose). To illustrate what the problem is with evolutionarily accounting for the truth of our thoughts Plantinga (in his 1994) quotes Patricia Churchland (1987): "Boiled down to essentials, a nervous system enables the organism to succeed in the four F's: feeding, fleeing, fighting and reproducing. The principle chore of nervous systems is to get the body parts where they should be in order that the organism may survive. ... Improvements in sensorimotor control confer an evolutionary advantage: a fancier style of representing is advantageous *so long as it is geared to the organism's way of life and enhances the organism's chances of survival* (Churchland's emphasis). Truth, whatever that is, definitely takes the hindmost." Churchland's point, according to Plantinga is that "Our having evolved and survived makes it likely that our cognitive faculties are reliable and our beliefs are for the most part true, only if it would be impossible or unlikely that creatures more or less like us should behave in fitness-enhancing ways but nonetheless hold mostly false beliefs." Plantinga argues that it is not unlikely at all, since even if we assume that beliefs *qua* beliefs are causally efficacious (i.e., that the cognitive content of mental states play a causal role in producing behaviour – an assumption that seems to be in contradiction with how Flew sees the relation between mental content and causation), it is not beliefs that do the causing, but belief-desire pairs, and the same adaptive behaviour can be caused by many different belief-desire pairs, many of them involving beliefs that are false. Plantinga considers imaginable cases when creatures like us behave along the same adaptive patterns as we do, although most of their beliefs are false. He concludes that the truth of mental content is very unlikely to be adaptive, there is no selection pressure on truth. I don't want to evaluate his argument here, since we can do without its

Maybe it is worth to sum up briefly where we are. Anscombe pointed out a hidden, unargued premise in Lewis's argument, namely that Lewis thinks that a rational explanation and a non-rational causal explanation for the same thought are competing explanations that cannot be true simultaneously. Anscombe and Flew argued that this hidden premise is false. Both the causal and the rational explanations are answers to the question "why is it that you think what you think?", but in the case of the two different answers the "why?" of the question is interpreted differently. The two answers allegedly answer two equivocated but entirely different questions. However, as it was argued, this view bears the consequence that reasons do not interfere with the causal process that brings it about that one thinks what he thinks. So there is a powerful objection to the Anscombe-Flew argument against Lewis's hidden premise, namely, that if rational and causal explanations do not interfere, and beliefs are effects of non-rational causes, then rationality does not empower us to resist beliefs we have no ground to believe, so it is entirely fortuitous if we believe what is rational to believe, and this is absurd. Flew suggested that there is an evolutionary explanation to why it is that a significant portion of what we believe is also rationally justified. In his view there is a harmony between non-rationally caused belief, on the one hand, and rationality that cannot interfere with causal processes, on the other, pre-established by natural selection. But this suggestion can be discarded, because if reasons are as idle relative to the causal order as Flew suggested, then natural selection is totally blind to such a harmony.

I conclude that this way of resisting Lewis's hidden premise is not viable.

The other way: maybe reasons are causes

But there is another way. The above discussion was based on the suggestion that the two explanations do not compete because they are answers to different (but equivocated) questions, and there is, of course, nothing wrong with having two different answers to two

---

conclusion. To tell the truth, I am not entirely convinced by what he is saying. I just wanted to note that the adaptiveness of the mere representative adequacy of our cognitive faculties can be reasonably drawn into question. (The part of Plantinga's 1994 from where the above quotes were taken is a summary of the argument given in the last chapter of his 1993 book *Warrant and Proper Function*.)

different questions. This line of thought broke down. But another solution suggests itself. Maybe there is no equivocation upon the "why", the question "why is it that you think what you think?" demands a unique answer, but the rational and the causal explanations aren't two different answers, only two descriptions of the same answer. The puzzlement of having one too many answers for a question is solved not by splitting the question but by uniting the answers. The idea is that effective reasons are causes, and when a reason is cited in a rational explanation, and when a neural occurrence is cited in a causal explanation for the same thought, we are referring to the same thing under different descriptions.

This is familiar.[142] But can it be right?

Prima facie contentful mental occurrences like reasons do not look at all to be related to each other or to physical events as if their contents were either effects or causes. The relations they bear to each other or to physical events seem too loose to be causal. There is a long tradition, a broadly Wittgensteinian one, that holds that, as far as there are rules to relate mental content to other occurrences (physical or mental), these rules are not descriptive truths that, when studied carefully enough by psychology, one time would be refined to take the shape of generally valid empirical laws, but norms, which can be either followed or neglected.

The categorial difference between generally valid empirical laws and norms is exemplified by the laws of logic, as it was classically exposed by Husserl and Frege arguing against the psychologism popular in their time. Consider two propositions, *p* and *q*, such that *p* entails *q*. Whoever has a justification to believe that *p* entails *q*, is entitled to deduce from it, and be justified in believing that non-*q* entails non-*p*. But empirically, the thought that *p* entails *q* is very often followed by the thought that non-*p* entails non-*q*, rather than the thought that non-*q* entails non-*p*, despite the fact that the latter follows logically and the former doesn't. Let A denote the thought that if *p*, then *q*. Let B1 denote the thought that if non-*p*, then non-*q*, B2 the thought that if non-*q*, then non-*p*. If the fact that A is followed by either B1 or B2 is due to a connectedness between the contents of the two thoughts, then this connectedness is hard to believe to be

---

142 As noted by Julia Tanney (1995), the wide acceptance of the thesis that reasons are causes, as proposed and defended by Donald Davidson from 1963 onwards, is sometimes referred to as one of the few achievements of contemporary analytic philosophy.

causal. If it is causal then how is it that the cause being the same content (A), the effect is sometimes B1, sometimes B2? If one attempts to account for the difference in the "effect" in terms of the mental history of the individual one finds that informedness about logic and dedication to think clearly reduces the frequency of B1's following A's dramatically. The most plausible explanation for this is that even if there is some sort of causal connectedness between thoughts with such contents, this is not in virtue of the exact contents of the thoughts involved, and the most significant factors that have a bearing on whether A is followed by B1 or B2 are the awareness of the applicable logical law, and the willingness to think logically, suggesting that what connects the exact contents is a norm, which can be followed, but can also be ignored or neglected. Norms cannot be construed as causal relations.

We may come to similar conclusions examining the connectedness of various propositional attitudes, not just beliefs, through more general "rules" of rationality, not just logical rules. A set of examples can be borrowed from Paul Churchland[143] arguing for his thesis that folk psychology is an explanatory theory.[144]

(i) (X fears that p) → (X desires that ~p)
(ii) ((X hopes that p) & (X finds out that p)) → (X is happy that p)
(iii) ((X desires that q) & (X believes that ($p \supset q$)) & (X knows he can bring it about that p)) → (X brings it about that p)

These are psychological generalizations. In (i) and (ii) propositional attitudes are connected, in (iii) propositional attitudes are connected with an action. The question is the nature of the relation represented by the arrow (→) in the three formulae.

The relation represented by the arrow may seem to be lawlike for some innocent substitutions of p and q, but turns out to be much less lawlike for substitutions about which our attitudes can be ambiguous, or in relation to which problems with the weakness of the will may

143 P. M. Churchland 1981.
144 Here and the next few paragraphs I am following Julia Tanney defending the thesis of the anomalism of the mental (Tanney 1995). Churchland's examples are used by Tanney.

arise. (Try to substitute "Y loves X", or "X is pregnant" for *p* in (i) or (ii).)[145]

Can these formulae be developed into generally applicable psychological laws by way of refining the propositional attitudes involved, or by adding additional conditions on the left hand side of the arrow, or by restricting the range of *p* and *q*, or even for some particular substitutions for *p* and *q*, by adding *ceteris paribus* clauses listing the exceptions? I think Tanney is right that this would be a hopeless enterprise, because what these efforts are likely to come to is "laws" of the form "If this-and-this is true of X, X does or thinks that-and-that, unless he doesn't". This is because of the categorial difference between descriptively valid laws and norms that can be followed or neglected, and because (i-iii) belong to the latter category.

But if this is right, then construing the relation connecting the content of thoughts (propositional attitudes) to physical occurrences that precede them (of which they are representations), or to actions or thoughts that they rationalize, or to thoughts that are rationalized by them, *as causal relations* is forging the relation too close, making the account unable to accommodate errors, i.e., cases of not following the norm.

One might suggest that one day psychology will discover true causal laws that are applicable descriptively, it is just that the mental occurrences figuring in those laws will not be individuated as propositional attitudes but some other way (as computational states, for example – as the Churchlands suppose). Then, arguably, this would be *the right way* to individuate mental states, and propositional attitudes would go like the phlogiston. But with propositional attitudes rationality would go as well, since it is them the norms of rationality connect with each other and with actions. It is very difficult to imagine that it will ever happen. To start with, I am clueless to what it would mean *to think that* propositional attitudes are unreal. It seems to be a contradiction in terms.[146]

What we are discussing now is an argument to the effect that the combination of physicalism, determinism, and the thesis that causes are mechanistic cannot be squared with the fact that we are rational. In this context the motivation behind the thesis that reasons are

---

[145] Tanney ibid. Anyone familiar with the phenomenon of the weakness of the will can construe an example when the relation represented by the arrow (whatever that may be) breaks down for (iii).

[146] See footnote 129.

causes is to offer a theory that accounts for how reasons can have a place, and a role in bringing about things, in a world which is a physical mechanism. So in this context this thesis is a particular case of the more general thesis that all mental occurrences are physical. According to this more general thesis reality is ontologically homogeneous and there are different levels of description to account for this homogeneous reality. The same entity can be identified as a contentful mental occurrence at one level of description (as a reason, for example), and as a physical occurrence at another level (as a physical cause or effect, for example).

What the suggestion that the two explanations, rational and causal, do not compete with each other, because they are the same, comes to is this: There are two thoughts A and B. A both rationalizes and causes B. Both A and B can be described at both levels, the mental and the physical. On the mental level A and B are described as the occurrences of contentful thoughts in the consciousness of the cognizer, defined by their content (most probably as propositional attitudes). On the physical level A and B are described as physically specifiable brain states. On the mental level A and B are connected by a rationalizing relation that holds between the two contents. On the physical level A and B are brain states of certain types that subsume a law of nature that prescribes that the occurrence of a brain state of the type of which A is an instance is to be followed by the occurrence of a brain state of the type of which B is an instance. The suggestion that reasons are causes is that these two are two descriptions of the same relation that holds between A and B.

But it cannot be so, if the above considerations are any good. Because the two relations are categorially different: one is conformity with a norm, the other is subsumption under a descriptively valid law. These two cannot be descriptions of one and the same relation.

Davidson endorsed the thesis that regularities at the mental level of description are norms and not descriptively applicable laws. In other words, he believed that there are no laws on the basis of which mental events predict, or can be predicted by, other events, mental or physical. (This is often referred to as "the anomalism of the mental".) Davidson proposed an ingenious solution to the problem how reasons can yet be causes. This is solution is known as "anomalous monism".

The idea is that the mental and the physical descriptions of the same event are not connected by bridge laws. The mental event is

identical with the physical event, this is a brute metaphysical fact. But the description of the content of the mental event that identifies it as a token of a mental type is not connected in a lawlike way with the physical description of the same event that identifies it as a token of a physical type. It is not the case that the mental and the physical descriptions would individuate the same type in two different levels of description. Metaphysical identity doesn't hold between the types, it holds only between the tokens.

It allows for the mental level of description to be anomalous, whereas a mental event (a reason) can be a cause in virtue of its brute metaphysical identity with a physical event, and in virtue of subsuming a causal law as a physical event.

But it doesn't seem to provide for what it takes to identify the causal explanation with the rational one. Reasons are causes on this account, that much is certain. It is in virtue of the brute metaphysical identity of the mental event (of having the reason) with the physical event (the cause). But there is no such metaphysical identity between the types to which the same event subsumes under its mental and its physical description. So there cannot be an identity between the rational and the causal *explanations* that refer to those types, respectively. So we have two different answers for the same question (why is it that you think what you think), after all, and it is not true that they are just two different ways of saying the same thing.

But maybe at this point the anti-Lewisian can switch back to Flew's position with a better hope to defend it. He might say that he found that the rational and causal explanations are entirely different, having nothing to do with each other, but now he can account for the difference between mere rationalizations and rational explanations that truthfully account of a belief's coming about. True rational explanations are those that identify the occurrence of the reasons in the cognizer's consciousness as the true rational grounds for the thought being explained, which are identical with the physical events that caused the physical event that is identical with the event of thinking the thought that is being explained.

But is it what we need? Does the rational explanation tell the truth about how the thought being explained came about? What seems to be the relevant story about that is the causal story. If anomalous monism is right, then the rational story is different one.

Can't both be true at the same time? Well, it seems that the reasons *as reasons* made no causal work at all in bringing about the

thought for which they are supposed to be the rational ground. All causal work is done by the physical properties of the events which are the events of thinking the thoughts which are the putative rational ground. The contents of the thoughts are causally idle. So the view must be that reasons are causes, but not in virtue of their content but in virtue of the brute metaphysical fact that they are identical with some physical events with some physical properties. But if mental types are not identical with physical types, and the mental properties are not related to the physical properties of the same event by bridge laws, then the content of the thought that is being explained and the content of the thoughts which are identified as the true rational ground for thinking the thought being explained have nothing to do with the rational explanation's being true.[147] And this is absurd.

If the relation between the mental and the physical levels is anomalous, and all the causal work is done at the physical level, then if the events that caused the thought being explained have a content that is connected to the content of the thought that is being explained in any way is a coincidence, a matter of pure luck.

Now, as the last refuge, may come the appeal to evolution to eliminate the blind fortuitousness of the rationality of some significant portion of what we think. But this we have already discussed. The rationality of our thought is as invisible to natural selection on the Davidsonian account as it was on the original version of Flew's.

So I think we may conclude that trying to show that Lewis's hidden premise is false this way is as hopeless as the other way was.

How could Davidson be wrong?

There is still a remote chance of saving Anscombe's claim that causal explanations and rational explanations are non-competing ones. This is if reasons are causes, are identical with physical events, but it is not true that all the causal work they do is done by their physical properties.

That all causal work is done by the physical properties of events is what followed from anomalous monism. For Davidson the motivation for adopting anomalous monism was that this was the only way to reconcile four theses that he endorsed. These four theses

[147] As it was argued by Jaegwon Kim, e.g. in his 1993.

are: 1) ontological physicalism, 2) that there is causal interaction between the mental and the physical, 3) that causation is nomological, and 4) the anomalism of the mental.

That anomalous monism is the only way to reconcile these four theses is easy to see. Monism is implied by physicalism, the first thesis. The remaining three exclude the possibility that the mental and the physical properties of the same event could be related in a nomological way: According to the second thesis there are mental events that cause and are caused by physical events. Is it possible that they cause, or are caused by, physical events, because their mental properties and the physical properties of the other event subsume a causal law? No, because then the fourth thesis would be violated. The only way not excluded by the fourth thesis for the mental event to cause or be caused is if this is because it is identical with a physical event, and its physical properties and the physical properties of the event which is caused by it, or which caused it, subsume one such law. Now suppose that the mental properties of these events are nomologically connected with the physical properties of the physical events with which they are identical (suppose that there are bridge laws that cover them). The physical properties of these events are nomologically connected with the physical properties of the events which are caused by them, or which caused them (there is a causal law that cover them), otherwise there would be no causal relation between the two. But then the combination of the causal law and the bridge law connects nomologically the mental properties of the mental event with the physical properties of the physical events which it caused, or caused it, which violates the fourth thesis. The only remaining alternative is if there are no bridge laws, i.e., anomalous monism. That satisfies all four theses. Physicalism holds, mental events are causally interacting with the physical realm in virtue of being metaphysically identical with physical events, and physical causation can be nomological without violating the anomalism of the mental due to the lack of bridge laws.

Do we have reasons to endorse the four Davidsonian theses?

Well, I think physicalism is false. But we are now in a *what-if* kind of argument, discussing what if physicalism holds.

One who endorses the thesis that reasons are causes (not just of other thoughts, but of actions, as well) cannot deny thesis 2. We are in a theoretical rather than a practical context now. We are discussing reasons that explain thoughts rather than actions. But some thoughts

explained will explain rational actions, and it is hard to imagine what would motivate a theory according to which reasons to believe are causes but reasons to act are not. Some of the thoughts that explain are presumably attributable to causal interactions with the environment. So, for the second thesis, the difference between the theoretical and the practical contexts is not relevant. If determinism is to hold it has to hold in every step from an environmental input to an action in response. So the mistake must be located in either of the last two theses.

I think we can exclude the fourth one. That mental content does not predict, and cannot be predicted, seems to be a very strong empirical fact. Mental types individuated as propositional attitudes notoriously resist subsumption under strict laws. Some say mental types are ill-identified as propositional attitudes. But it is very hard to make sense of this suggestion.

So the only remaining candidate is the third one. We are in a *what-if* type argument, and the clause after the *if* contains determinism. In chapter 2 I argued that a causal determinist should adopt the nomological account of causation, since this is the only option that would give him resources to make his thesis empirically plausible. But in the discussion of the present chapter nothing depends on whether determinism can be made plausible. It is concerned with whether determinism (in combination with physicalism) can be true, and it can be true even if it can't be proven or made plausible. So the argument I gave in chapter 2 for favouring the nomological account of causation will not do here.

But it is possible to avoid arguing about the nature of causation even so.

Consider what philosophical work denying the nomologicality of causation would do for the defender of the thesis that reasons are causes but not only in the odd Davidsonian way which rendered their description as reasons causally idle. Surely, whoever wants to see reasons *qua* reasons causally active should want to deny the nomologicality of causation to make it possible that the mental properties of the event of a reason's occurring in one's consciousness do some causal work, even thought the mental is anomalous. If causation doesn't have to be lawlike, then mental contents can cause and be caused without ceasing to be unpredicting and unpredictable.

Note that in this case it would be very hard to identify and check the truth conditions for a causal explanation, and a causal explanation

would hardly *explain* in the sense that it would make the sequence of events intelligible. Nevertheless, the causal explanation could be true.

But if the anomalousness of the account is to be retained, then mental properties should not be identical with physical properties. But then, if mental properties are allowed to do causal work, then the principle of the causal closure of physics should be abandoned. I don't think it is a price a physicalist is willing to pay for letting mental contents cause. Of course, this wouldn't amount to admitting entities or events which are genuinely non-physical, external to the physical realm, metaphysically speaking, to interfere with what is physical (interactive substance dualism). But it would break the dogma of the causal completeness of the science that is called physics, the description that can be given of the world at the physical level. The belief in this completeness is the stronghold of physicalism which, I think, most physicalists would defend to the last bullet, since this is the belief on which rests the other belief that the mental and the physical phenomena constitute just two levels of description for the same ontologically homogeneous reality.

So I don't think this is a real alternative for a physicalist.

The conclusions of the argument

The suggestions that reasons explanation and causal explanation do not stand in each other's way, because (a) they have nothing to do with each other, or because (b) they are the same, have been discussed, and I cannot think of any other ways to substantiate the Anscombian claim that they do not compete as explanations. Now it is time to review the philosophical price the "naturalist" (in Lewis's sense) has to pay if he is to maintain that Lewis's hidden premise, i.e., that the two explanations do compete, is false. Our discussion has shown that the naturalist has to endorse at least one of the following claims to resist Lewis's premise:

(i) We are immensely lucky to have rational grounds for a good deal of what we are caused to believe, and this luck cannot be given an evolutionary or any other explanation, so it is a mystery.

(ii) Despite all experience pointing to the contrary, propositional attitudes are not anomalous: they predict and can be predicted, and reduce to physical types.

(iii) Types of mental content are not really individuated as propositional attitudes: propositional attitudes are unreal.

(iv) Causation is not nomological, and physics is causally incomplete.

If the naturalist is reluctant to endorse any of these positions, this is not the end for him. It means only that Lewis's argument is sound and he has to choose between abandoning naturalism or abandoning rationality. So the naturalist has a fifth way to hold on to naturalism, and this is to embrace the position that

(v) Rationality is unreal: it is never really the case that we believe something because it is rational to believe it, and we don't really have the power to resist believing anything we are caused to believe even if it is irrational, and there is no principled reason why something that we are caused to believe wouldn't be irrational.

Of course, in this latter case, this applies to the very thought of naturalism itself, just like to any other thought. Naturalism can be true, but cannot be held rationally.

So, if our argument is right, the naturalist has to choose at least one from the menu of (i-v). Refusing to endorse at least one of them logically binds him to give up naturalism. I suspect that very few naturalists would choose (v). It may be an attractive option to deny that causation is nomological, but that is not enough in itself. And the other conjunct of (iv), giving up the completeness of physics, would deter most naturalists, I believe. (i) is not a real option. Naturalists do not like mysteries, and this is a really bad one. So, I suspect, most naturalists would opt for either (ii) or (iii). (iii) is actually endorsed by some naturalists (the eliminative materialists, like the Churchlands, for example), but this position seems to me very hard to make any sense of, at all. If there ever was a self-defeating position, then this is.[148] So (ii), denying the anomalousness of the mental is the winning option.[149]

---

[148] For a discussion of whether eliminativism about propositional attitude is literally self-refuting or just immensely implausible see Boghossian 1990 and Devitt 1990.

[149] For a physicalist who sees the implausibility of a type-identity thesis, and so is inclined to anti-reductionism about mental content, but seeing the alternatives, considers giving up his anti-reductionism and embracing position (ii) rather than any other of the options see Paul Boghossian commenting on this situation in a passage quoted by Tanney (Boghossian 1989): "Finally, though, there is the question of mental causation: how are we to reconcile anti-reductionism about mental properties with a satisfying conception of their causal efficacy? It is a view long associated with Wittgenstein himself, of course, that propositional attitude explanations are not causal explanations. But, whether or not the view was Wittgenstein's, it has justifiably few adherents today. As Davidson showed,

But if philosophy is to be informed by what we learn about the world and ourselves empirically, rather then by pre-chosen philosophical commitments, then we have to consider that if the choice is between the anomalism of mental content and naturalism (in Lewis's sense, that is, the conjunction of ontological physicalism, determinism, and the thesis that causation is mechanistic), we have much more empirical evidence for the former then for the latter.

So we may conclude that, quite probably (unless we are fundamentally misinformed empirically in respect of the anomalism of the mental), the falsity of naturalism (in Lewis's sense) is a precondition for rationality. Then, quite probably, naturalism is false.

What if we drop determinism?

Lewis's argument was a variation on an argument originally offered by Epicurus against determinism. We have seen that the argument probably works against what Lewis called "naturalism", which is a conjunction of determinism with physicalism and the thesis that causes know no purpose or meaning. What changes if we drop determinism from this conjunction? The conjunction that Lewis called "naturalism" is not necessarily held by present day naturalists. Although there are deterministic interpretations to quantum mechanics (see the Appendix[150]), for all we presently know, the physical realm may well be objectively indeterministic, and, although it is not meant to be a philosophically respectable argument, probably the majority of the scientific community thinks it is. So it is interesting to see what changes in the argument, if, on behalf of the naturalist, we allow for objectively indeterministic physical events. Is the probable incompatibility we pointed out between rationality and "naturalism" dependent on determinism being part of the latter?

I think not. Lewis's problem with what he called naturalism was that on that doctrine thoughts were events that, just like any other event, were the products of non-rational causal processes. If we drop

---

if propositional attitude explanations are to rationalize behaviour at all, then they must do so by causing it.... But...how is an anti-reductionist about content properties to accord them a genuine causal role without committing himself, implausibly, to the essential incompleteness of physics? This, I believe, is the single greatest difficulty for an anti-reductionist conception of content. It may be clear that it will eventually prove its undoing. But the subject is relatively unexplored, and much work remains to be done."

[150] The Appendix figures only in the longer version of the thesis to be found on the internet.

determinism, the only change is that a thought is not necessarily determined by such a process. It may be undetermined by its causes. If we are in a quantum mechanical world, then most probably in any given occasion there is a set of possible thoughts, and the extension of the set is determined by physical causes, plus physical causes assign probabilities to each member of the set, and that is all. Now if this is all to be said about how this world is different from the original one, in which there was no place for rationality, then there is no place for rationality in this one either. It turned out that a rational thought cannot arise as a result of a deterministic physical causal process. It seems clear that a causal process with some indeterminacy injected into it cannot do any better.

Some might object that this is too quick. When I argued against one of the possible objections to the Lewisian argument, the objection that rational and causal explanations do not compete because they are one and the same, I appealed to a categorial difference between rational processes and causal ones on the ground that the latter follow strict descriptive laws, whereas the former are anomalous. I also claimed that the anomalism of the mental was an empirical fact. Now if we let causal processes accommodate randomness then, perhaps, there is not so much of a difference any more between how we observe ourselves thinking and acting, and what we would accept from causal, that is, somewhat random yet mechanical (say, quantum mechanical), cognizers and deliberators.

I have two considerations to offer in response.

If rational thinking reduced to a physical mechanism involving quantum mechanical indeterminacies that propagate to the macro level (supposing the quantum mechanics is genuinely indeterministic), the statistical laws of quantum mechanics would nevertheless apply to it, so in this sense it would not be anomalous. However, rational creatures do not seem to follow in their ways of thinking and acting statistical laws like those of quantum mechanics. If they did, in large numbers we would constitute highly predictable communities, like the pan-galactical population in Isaac Asimov's famous *Foundation*, whose future was quite precisely calculable for a bunch of qualified psychologist-mathematicians, the so-called psycho-historians, as affects on the course of the history of the Galaxy due to individual deviations from the mean cancelled out (at least until a very powerful and very deviant individual, the Mule, came along). We just don't seem to work like that. Politicians often use thumb-rules for

predicting how the public would react to what they do and say, but at least as often as not, they get it wrong. If mathematical-statistical modelling of the behaviour of rational creatures was possible, it would be a powerful industry by now as it would obviously attract the attention of mighty investors. But nothing like that is anywhere near.

But, more importantly, it just doesn't seem right to reduce rational thought to a mechanism involving randomness. Such a model could not account for the normative character of rationality. It is hard to see how a thought that popped up genuinely at random would constitute a rational thought.

Now maybe rationality is the quality of the thoughts that the randomness-involving mechanism would produce if it did not involve randomness, and the randomness in the mechanism accounts for the empirical anomalism of our thinking, as deviations from rationality due to random perturbations of an otherwise deterministic mechanism, which is both physical and mental, i.e. the mental properties of the events they consist of reduce to their physical ones.

It doesn't seem to work either. When we discussed cases of practical deliberation embedded in theoretical thinking a little earlier we have seen that in many cases there is not a uniquely right option in such situations (for example in the case of terminating the checking process for an empirical generalization that is to be used later as a general premise in a deductive argument), yet it would be destructive for intellectual responsibility, and so for rationality, if we had no genuine choice from the options because the 'choice' is made for us by a random element of our psycho-physical mechanism.

So the conclusion seems to be that the event of thinking a thought because of having a rational ground for it cannot be identical with a physically determined event, nor can it be a random event. So there must be events which are neither determined by physical causes, nor random, for their being rational thoughts.

If we are to maintain the thought that everything that is non-random is caused by something, then rationality requires at least property dualism, with mental properties having causal powers independently of the physical properties that are co-instantiated with them. It is very hard to imagine these causal powers without effects in the physical world. As long as ontological physicalism is maintained, mental properties have to be instantiated by entities that have also physical properties. So, if a mental cause has a mental effect, it has to have a physical effect, too: the obtaining of the physical properties

that are co-instantiated by the mental effect. So, even if the physical realm is allowed to be indeterministic, rationality requires at least interactive property dualism. (To the same conclusion we may come by extending our discussion to rationally explaining actions, not just thoughts, without having to appeal to the co-instantiation of mental and physical properties.)

What if we drop physicalism?

But can rational inference be causation by irreducibly mental properties? Is there room for rationality in a deterministic world, whether or not it is physicalistic? Now we are back with the original Epicurean question. It can be answered by examining what changes in the argument if we drop physicalism, instead of determinism, from the conjunct the Lewisian argument has shown to be incompatible with rationality.

The argument was concerned with whether there is room for a truthful rational explanation (not just a mere rationalization) alongside a full physical causal explanation, or, alternatively, whether the two can be the same thing under different names. In the first part of the argument, when we came to the conclusion that a rational explanation cannot be more than a mere *post facto* rationalization if there is an independent causal explanation we made no mention of whether the causal explanation was physical or not. In the second part of the argument, when we discussed whether reasons can be identical with physical causes, we came to the conclusion that there was a categorial difference between rational and causal relations. And when we argued for that, the argument never traded on the causal relations' being relations between physical causes and effects either. Whether or not causal relations are between physical causes and effects was indifferent, the categorial incompatibility was not dependent on that.

The categorial difference pointed out between rational and causal explanations was, however, dependent on the truth of two principles, 1) the anomalism of the mental, and 2) the nomological character of causation. The anomalism of the mental is a strongly confirmed empirical principle. The discussion we gave to this principle was not dependent on the truth or falsity physicalism, so it can be retained if we drop physicalism from the position against which we are arguing. But we never *really* argued for the nomological character of causation. We just noted that dropping this principle, and letting the anomalous

mental realm do causal work is not a real option for our opponent, because that would amount to the abandonment of the causal completeness of physics, which he probably wants to retain. But this manoeuvre was dependent on our opponent's being a physicalist. If physicalism is dropped from the position we are attacking, we cannot argue this way.

I really believe that causation is nomological. But I cannot argue for it here, this is a vast topic.

So the responsible thing from my part is to say that the conclusions I draw from this point on should be read with keeping in mind that they are dependent on the nomologicality of causation, for which I haven't argued.

But let me mention that I think there is a deeper trait in causation's character that is only manifested in its nomologicality but not identical with it, and that it is because of this deeper fact that rational explanation must be categorially different from causal explanation. This deeper fact is the mechanical nature of causation, which, I think, is endorsed even by those metaphysicians who believe in non-nomological causation.

Lewis's problem with causation by physical causes was that he thought it must have been a non-rational process, so it couldn't produce a thought whose coming about is truly accounted for by a rational explanation. When we argued in defence of his position, we did so by pointing out that physical causes cannot have a rationalizing content, because propositional attitudes cannot be identical with physical causes. But this is just one of the reasons why a physical causal process must be non-rational. The other reason is the mechanistic nature of causation. The problem is not just that "the causes are non-rational", but that the way the process goes is non-rational too. This latter reason to hold causal processes non-rational remains even if causes are thought to be mental contents. Even if reasons could be causes, anomalous causes, it is far from clear that causation by mental content would be a "rational process", as long as causation is thought to be mechanistic. Seeing that A is a rational ground for B, and thinking that B *because of that* just doesn't seem to be the same thing as the belief in B being mechanically produced by belief A, whether or not that mechanical production is attributable to the perfectly mental content properties of the two beliefs in question. Maybe the emergence of thought B in the mind, and the ignition of the checking process which is to determine whether the cognizer

finds it that A is a firm enough rational ground for B can be given a mechanistic model in which every step is determined by forces inherent in what already there is in the mind. But the judgement, the termination of the checking process, seems to be something that resists subsumption to such a model. Or else, it ceases to be a judgement (performed actively, freely, on the initiative of the cognizer), and the whole process ceases to be rational, or more precisely, ceases to be one to which either of the rational/irrational predicate-pair can be applied. The element of judgement, and the fact that a judgement cannot be identical with a step in a mechanistic process, is what is responsible for the anomalism of mental content. The incompatibility between judgement and mechanism is deeper than the incompatibility between the anomalism of the mental and the nomologicality of causation. The former incompatibility remains even if we give up the thought that causal relations must be lawlike.

I admit that this is not a very solid argument. It is much more like the statement of the intuition that a judgement cannot be a passion, and everything caused mechanistically by causes (which were none of our making) is a passion. I think it is a very strong intuition, about which I will say more in the eighth chapter. But it is important to keep in mind that, in order to maintain that rationality is incompatible with determinism, I have to rely on this intuition only if doubts arise concerning either the anomalism of the mental or the nomologicality of causation.

How is rationality possible, after all?

I think we have given strong arguments in support of Lewis to the effect that if the world is a deterministic physical mechanism, then there is no place for rationality in it. We have seen that nothing really changes if we drop either determinism or physicalism or both. The final verdict, I believe, is that if the world is a causal mechanism, in which every event is causally determined, or, to the extent it is causally underdetermined, random, then there is no place for rationality in it. (If my last argument about the incompatibility of rationality and the mechanistic character of causation is deemed unconvincing, then add: "provided that causation is nomological"[151].) So the Lewisian argument is, at bottom, not just an argument against

[151] The anomalism of the mental is so strongly confirmed empirically that it would be overcautious to add a clause about that every time.

physicalism, nor is it just an argument against determinism, but it is an argument against the thesis that there are only causally determined and random events.

So it seems that the only thing that can save rationality is libertarian freedom, which makes it possible that there be things that are neither caused nor random. If there are rational thoughts at all, they must be judgements made in a libertarian free way (provided that causation is nomological).

Can we account for the difference between the reasons that were really effective and those that can merely serve the purposes of post facto rationalization in these terms?

I have argued that reasons explanation and causal explanation are categorially different. Our problem with Flew's suggestion that reasons explanation and causal explanation are answers to two completely different "why?'s" was that on this ground it seemed impossible to distinguish between rational explanations that tell the truth about how a propositional attitude came about, and mere post facto rationalizations. Davidson thought that this distinction cannot be accounted for unless we construe the relation between the *explanans* and the *explanandum* causally. Aren't we falling back to Flew's position if we reject the idea that reasons are causes?

No. The reason that truthfully explains is the one that the cognizer considered and judged to be a firm enough ground to adopt the propositional attitude that is being explained. I suggested that the judgement should be construed as a libertarian free mental act. A libertarian free mental act is not caused, so there is no unbroken chain of causal determination leading from the *explanans* to the *explanandum*. Nevertheless, the *explanandum* comes into being in the very act of performing the judgement. Performing the judgement on how one should relate himself to a proposition and adopting a propositional attitude are the same. So if there was a causal explanation, independent of the judgement based on the reason that truly explains, for the adoption of the propositional attitude being explained, then there would be no room for the judgement to be performed in the libertarian free way. So on the account I am proposing, it is true that the *explanandum* comes into being in a non-causal way, nevertheless, it comes about in a way that interferes with the causal order in the sense that it does not tolerate an independent causal explanation.

I think the predicate "the reason on which the judgement resulting in the adoption of the propositional attitude was performed" clearly identifies the reason that truly explains and distinguishes it from other reasons that can merely serve the purpose of a *post facto* rationalization. It is just that the relation that is thus identified between the reason that truly explains and the propositional attitude that is being explained cannot be specified in an impersonal language. But that shouldn't be a problem as long as we don't feel obliged to keep with physicalism.

More should be said about this in the context of explaining actions rather than thoughts in the eighth chapter.

Conclusions for physicalism

In this chapter we examined the conditions that need to be fulfilled for there being rational thought. So the arguments of this chapter were "transcendental arguments" in the Kantian sense. One of the important conclusions of the argument is that some sort of dualism seems to be a precondition for rational thought.

Rational thought means that there are propositional attitudes that are adopted because it is rational to adopt them. Adding on an argument offered by C. S. Lewis we have seen that the adoption of a propositional attitude on a rational ground cannot be identical with a physical event that is causally necessitated by other physical events, and an underdetermined, and to that extent random, physical event is not a better candidate either. The facts that make a thought rational must be external to the physical realm. There is still the possibility that those facts are irreducible mental properties of some events that coinstantiate physical properties as well. This would save physicalism in the sense of ontological monism in combination with property dualism.

Up to this point the argument seemed conclusive. So rationality requires *at least* property dualism – and, as our analysis has shown, that property dualism must be interactive, contradicting the idea of the causal completeness of physics. (The analysis of the suggestion that rational thoughts are caused by causally efficacious mental content showed that this way of accounting for rationality is probably not viable either, and the likely conclusion is that rationality cannot be given an impersonal account at all, so it requires dualism also in the

sense that the personal account of some things must be considered irreducibly personal.)

But isn't the causal closure of physics an empirical fact?

In an article titled "The Rise of Physicalism"[152] David Papineau argued that "the scientific evidence" for the causal closure of physics was the main reason why physicalism became the dominant position in the philosophy of mind in the second half of the last century. The argument from causal closure is the Master Argument for physicalism according to Papineau, without which physicalism could have never reached the status it has by now.

As far as this last point is concerned, I agree. Without the argument from causal closure physicalism is a philosophical policy badly lacking an argument to underpin it.

But when it comes to explaining why we should believe in the closure hypothesis Papineau simply says it is not a respectable thing to do by a philosopher to doubt it. He writes:

> Of course, as with all empirical matters, there is nothing certain here. There is no knock-down argument for the completeness of physics. You could in principle accept the rest of modern physical theory, and yet continue to insist on special mental forces, which operate in as yet undetected ways in the interstices of intelligent brains. *And indeed there do exist bitter-enders of just this kind, who continue to hold out for special mental causes, even after another half-century of ever more detailed molecular biology has been added to the inductive evidence which initially created a scientific consensus on completeness in the 1950s.* Perhaps it is this possibility which Stephen Clark has in mind when he doubts whether any empirical considerations can "disprove" mind-body dualism. If so, there is no more I can do to persuade him of the completeness of physics. *However, I see no virtue in philosophers refusing to accept a premise which, by any normal inductive standards, has been fully established by over a century of empirical research.*[153]

---

[152] Papineau 2001.

[153] Papineau 2001, last paragraph, emphasis mine. (Note that if the arguments of this chapter are correct, there is no need for special mental causes for rationality, as rationality turned out to work non-causally.)

Papineau's arrogance is not exceptional. But as compared to this arrogance the explanation he gives of how the closure hypothesis is supported by science is poor.

The explanation is given in the Appendix of his 2002 book titled *Thinking about Consciousness*, and can be reconstructed as follows: (1) The four known forces respect the conservation laws. (2) For all we know, they are universal, and there is no empirical sign of the existence of any other (non-conservative) force. (3) Conservative forces make physics causally complete. Therefore, (4) physics is causally complete.[154]

Now the idea that conservation laws make physics causally complete is classic one. Leibniz, for example, thought that what made it possible for Descartes to be an interactive dualist was his ignorance of a conservation law, namely the conservation of momentum. Descartes, according to Leibniz, was aware of the conservation of the quantity of motion, i.e. scalar momentum, but failed to take notice of the conservation of the direction of motion. The latter puts additional constraints on the movement of bodies. While Cartesian dualist interaction is consistent with the conservation of scalar momentum, since the direction of the motion can be altered without any change in the total quantity of motion, it violates the conservation of vectorial momentum, which prescribes that the direction of the motion should also be preserved throughout physical processes, leaving no room for the irreducibly mental to make any difference to the course of the movements of atoms by which both Leibniz and Descartes thought our brains were constituted. This was the motivation behind the theory of the pre-established harmony between the mind and the body: saving the genuinely mental without letting it causally interact with the physical, since interaction is ruled out by the conservation principle (Leibniz 1952).

Barbara Montero reviewed a long line of authorities following Leibniz holding the view that conservation laws entail the falsity of interactive dualism through the causal closure of physics.[155]

---

154 The same line of reasoning is given is his 2001.

155 Montero 2006. She quotes Dennett 1991, p. 35, Fodor 1994, p. 25, van Inwagen 2002, p. 196, and Crane 2002, p.48: "[A]ccording to Daniel Dennett *»[the] principle of conservation of energy…is apparently violated by dualism«* and that *»this confrontation between quite standard physics and dualism….is widely regarded as the inescapable and fatal flaw of dualism«*. Jerry Fodor expresses the same view: *»how«* he asks, *»can the nonphysical give rise to the physical without violating the laws of the conservation of mass, of energy and of momentum?«* In Peter van Inwagen's

But this argument is plainly wrong. For physical systems with sufficiently many degrees of freedom all conservation laws taken together fall short of prescribing a unique trajectory along which the evolution of the physical state of the system would be to continue. They are simply too few equations for two many variables. The mind could interfere in the evolution of the physical state of the brain without getting into conflict with any of the conservation laws.

It is the classical dynamical description of the evolution of physical systems that holds out hope to establish physics' causal closure. The method of Cauchy problems, to be discussed in the next chapter, is paradigmatically deterministic and therefore causally closed. A scientifically informed argument for causal closure should appeal to dynamical laws obtained in the form of Cauchy problems rather than conservation laws.

However, deterministic dynamical laws would exclude interactive dualism only if they were true of systems in which conscious minds are present. There is no proof yet that they are.

Moreover, the rise of quantum mechanics made it probable that the dynamical evolution of the physical realm is not deterministic after all. Papineau thinks that he can deal with this problem in a footnote.[156] He thinks that even if quantum mechanics is objectively indeterministic its probabilistic laws exclude interactive dualism, because any intervention from the irreducibly mental would contradict its probabilistic predictions. In the next chapter I will show that this argument is fallacious.[157]

So crudely put, determinism (verified for systems including minds) is vital for the hypothesis of the causal closure of physics to have any grounding in solid science. So if determinism and rationality exclude each other, this is bad news for the alleged scientific grounding of the closure hypothesis as well.

Conclusions for determinism

---

words, interactive dualism seems to *»require a violation of well-established physical conservation laws like the law of the conservation of energy.«* While Time Crane states that *»mental causation would…have to introduce 'more energy' into the physical world, thus violating the conservation laws«* ". It could be added that Herbert Feigl who advocated an early version of the knowledge argument also thought that the conservation laws entailed the falsity of dualism (1958).

[156] 2001, footnote 2.

[157] About the question whether quantum mechanics is objectively indeterministic please refer to the Appendix.

We started off our discussion with the Epicurean intuition that determinism renders a cognizer passive in respect of everything he thinks, and it is incompatible with knowledge, since knowledge is a propositional attitude which has an active character. Knowledge is not a passion. If the determinist is right, then everything is a passion. So either determinism is wrong, or no thought constitutes knowledge (including the very thought of determinism, so being a determinist is a self-defeating venture).

First we disentangled our case from the externalism-internalism debate in epistemology by pointing out that what we are really interested in is whether we can be related to the truth as we normally, commonsensically think we are—among others, we want to know whether we can be committed to the truth of a proposition in an intellectually responsible way, like when we think we know it in virtue of having a firm rational justification for it. Whether there are respectable accounts of knowledge that make no reference to justification is not relevant for our concerns, for when they identify knowledge without mentioning justification they surely do not identify a type of propositional attitudes that is individuated by a specific sort of (conscious and transparent) relatedness to propositions. The real Epicurean concern is not whether the determinist can be said to know that determinism is true on some respectable externalist account of knowledge, but whether he can be related to the claim he is making in an intellectually responsible way that qualifies him for being taken seriously in a debate.

Then we made a catalogue of the elements of rational justification that can make our commitment to the truth of some propositions intellectually responsible. Building on Lewis we examined one element of justification: rational inference, and found that inferring a thought from others that serve as a firm rational ground for it cannot be identical with an event necessitated by physical causes. (Unless we are fatally wrong about the anomalism of mental content.)

We have examined the possibility that it is an event necessitated by irreducibly mental causes, and found that it cannot be, as long as we hold on to the nomologicality of causation. And even if we consider the nomologicality of causation to be reasonably debatable, the mechanistic character of causation would render every thought caused by mental causes a passion, and the intuition that rational inference is not a passion is strong. But appeal to this bare intuition is necessary only if doubts arise concerning the anomalism of the

mental or the nomologicality of causation. Otherwise the incompatibility of rational inference with determinism seems to be conclusively proven.

Some conclusions for the whole project

I consider the possibilities that we are wrong about the anomalism of the mental and that there can be anomalous causation very remote. (But it should be kept in mind that I haven't argued for the latter of these claims.) So my conclusion is that determinism is very probably incompatible with rationality. So determinism is very probably false.

I have argued in the second chapter (and I will add to it in the next one) that there is not much principled reason to believe in determinism unless one thinks physics is deterministic, and physics is causally closed, so everything with causal powers to affect the physical realm, must itself be physical, and so everything with causal powers to affect the physical realm must evolve deterministically. I have also argued that the alleged evidence that physics is causally closed and deterministic is dependent on a nomological account of causation.

I think the arguments of this chapter made serious damage to this way of arguing for determinism. Unless we are dead wrong in holding on to the empirically very strongly supported anomalous image of mental content, the nomological nature of causation and the causal completeness of physics surely cannot be maintained together, or else, there is no rationality.

If my arguments in this chapter and in chapter 2 (and 6) are good, it is nearly the end for determinism. I think the empirical evidence for the anomalism of the mental is strong. In the next chapter I will argue that the alleged empirical evidence for determinism is very weak. So if the choice is between the anomalism of the mental and determinism, on the normal standard of rational thought and empirical enquiry, we have every reason to prefer the anomalism of the mental.

These findings conclude the first larger unit of my thesis. In this unit my aim was to show that we have every reason to be dissatisfied with the causal conception of control and freedom as the shallowness of this conception of freedom and the lack of genuine alternatives are really damaging to the values we naturally associate with freedom: self-determination, moral responsibility, rationality and intellectual responsibility. So, contrary to the bold claims of Dennett for example, it is a philosophically very well motivated programme to try to defend

the libertarian conception of control and freedom against the charges that it is empirically impossible, or that it is an incoherent idea, or that it necessarily reduces freedom to irrationality. I think it should be clear now that we have a lot to lose if the libertarian conception of freedom comes out untenable.

Whether it is is the subject of the remaining chapters. First I will deal with the empirical issues, and then I will turn to the question of the coherence and the rationality of the libertarian conception of control.

6 Can We Have Alternatives Anyway? – A. A Note on Determinism

Many philosophers treat determinism as an empirical fact, or a hypothesis that has been so strongly corroborated that no theory of freedom that is incompatible with it is worth much attention. In my view determinism is scientifically unfounded. In order to spare space I present my argument against the view that determinism is a near fact in a very condensed form.

Psychological determinism

Roughly two and a half centuries ago David Hume declared that facts of desires, hopes, worries, doubts, beliefs and knowledge, and those of intentions to act relate to each other as causes and effects. He claimed that there is a constant conjunction between motives and action, and that we know it empirically, and that this is so evident that no one who knows what he is talking about is ever expected to dispute it. He wrote:

> [I]t appears, not only that the conjunction between motives and voluntary actions is as regular and uniform as that between cause and effect in any part of nature; but also that this regular conjunction has been universally acknowledged among mankind, and has never been the subject of dispute, either in philosophy or common life.[158]

He said that those who deny it deny it only in words but their conduct shows that they in fact accept it. We are all largely dependent on the co-operation of others and we all rely on the necessary connection we assume to exist between the motives and the behaviour of the others in the strategies we follow in our lives. Hume thought that the same reliance on the connection between motives and behaviour is manifested in our social institutions of reward and punishment. The connections between motives and behaviour have

---

[158] *An Enquiry Concerning Human Understanding.* Section 8, Part I, 69. Hume 1975, p. 88.

been experienced many times, and our practices show that we trust firmly that they remain as they were experienced:

> The mutual dependence of men is so great in all societies that scarce any human action is entirely complete on itself, or is performed without some reference to the actions of others, which are requisite to make it answer fully to the intention of the agent.... In all those conclusions they take their measures from past experience, in the same manner as in their reasonings concerning external objects; and firmly believe that men, as well as the elements, are to continue, in their operations, the same way they have ever found them.

And:

> All laws being founded on rewards and punishments, it is supposed as a fundamental principle, that these motives have a regular and uniform influence on the mind, and both produce the good and prevent the evil actions. We may give to this influence what name we please; but, as it is usually conjoined with the action, it must be esteemed a *cause,* and be looked upon as an instance of that necessity, which we would here establish.[159]

Hume was of course aware of the fact that people often behaved differently from what was expected. He accounted for this phenomenon by claiming that whenever our predictions of the behaviour of others fail, it is due to the fact that our knowledge of their motives was not detailed and accurate enough. In other words, if there is an apparent indeterminacy in the link between motives and action, it is only epistemic.

What Hume suggests is that if a conjunction of the form "If this-and-this is true of X's motivational state, X does that-and-that" fails to be constant, then there is always another conjunction of the form "If this-and-this is true of X's motivational state, X does A, unless condition C is also true of X, in which case he does A' instead", which is constant. The *ceteris paribus* clause can be built into the description of X's motivational state, resulting in a more detailed and

---

[159] Ibid. Section 8, Part I., 76, pp. 97-8.

accurate description of it, and with this description we get a truly valid generalization of the original format.

It may be so. But this, of course, is not an empirical fact. This is only a hypothesis offered in account of the empirical fact that human action loosely conforms expectations based on known or assumed motives.

Two and a half centuries of empirical psychology since Hume's time failed to verify this hypothesis. Truly general empirical generalizations of this format are nowhere near. As Julia Tanney put it, what this business of amending psychological generalizations that failed to be truly general by adding such clauses may eventually come to is 'laws' of the format "if this-and-this is true of X, X does that-and-that, unless he doesn't".[160]

There is an alternative explanation for the empirical facts to which Hume appealed, and that is that motives strongly affect behaviour but fall short of determining them, and that there is a categorial difference between the normative laws of rationality and the descriptive laws of causation, as we have discussed it in the previous chapter.

Determinism from below

If the determinism of the mind cannot be established empirically on the ground of facts observed at the explicitly mental level, it is still possible to argue for determinism "from below". Determinism from below is a conjunction of two theses, a) that the mental events whose determinatedness is in question are redescribable at a "deeper" level of description, e.g., as neural events or physical events, and b) that there is strong scientific evidence that the evolution of that deeper level (neural or physical) is deterministic. This is an ancient idea whose roots go back to presocratic natural philosophy. Only the description of the underlying realm to which the mental realm is reduced, and the alleged scientific evidence for its determinism is new.

So to have a strong objection against libertarian freedom one needs to have a strong argument for psycho-physical, or psycho-neural reductionism, and a strong argument for the determinism of physics or that of the evolution of our neural states.

---

[160] Tanney 1995, see the previous chapter.

As it has been already pointed out, the two issues are not independent of each other. From the determinism of the underlying realm the reducibility of the part of the mental realm that plays a role in action production to that realm follows, as long as it is secured that an action has a description as an event of the underlying realm. It is uncontroversial that the idea of freedom involves freedom in respect of actions that have neural or physical aspects.

To see that reductionism is not dialectically independent of determinism we have to give a precise definition to determinism at these levels.

In the second chapter I gave the following definition to physical determinism:

The set of all physical events (U), of which the set of actions is a subset (A), has the property that there are core subsets within U, such that with a core subset and with the laws of physics only one totality of U is logically coherent, therefore, only one subset A is coherent; and the set of all events which are past relative to any arbitrarily chosen moment of time in any arbitrarily chosen frame of reference is such a core subset.

The analogous definition of neural determinism is:

The set of all neural events of an agent (N), of which the set of (the initiation of) all his actions is a subset (A), has the property that there are core subsets within N, such that with a core subset, with the laws of neuroscience, with a given input from the sense organs, and with a given totality of other physical influences coming from outside the neural system, the latter two treated as a fixed set of boundary conditions, only one totality of N is logically coherent, therefore, only one subset A is coherent; and the set of all events which are past relative to any arbitrarily chosen moment of time is such a core subset.

In the second chapter I argued that even if there may be a deeper account of causation than the nomological account – which has to be assumed if these definitions are to capture *causal* determinism – the evidence to which determinists can appeal are all on the nomological level, there is no evidence for determinism that would invoke empirical data about causation in a sense deeper than the nomological sense.

So in the second chapter I suggested to use the adjectives "causal" and "nomological" interchangeably for the purposes of our discussion of the question whether determinism can be considered as a scientifically well-grounded thesis.

If the above definitions of determinism are accepted, the causal closure of the physical realm follows from physical determinism, and

the causal closure of the neural realm, apart from the physical input provided by the sense organs and other external physical influences, follows from neural determinism. If the physical realm is deterministic under the above definition, then it allows for no interference from non-physical mental events or states of affairs to make a difference to its evolution. The same is true of the neural realm.

If the causal closure of an "underlying" realm that is evidently involved in actions is established, then the part of the mental realm that is involved in action production must reduce to it, otherwise it could not have any role in bringing about actions.[161]

Therefore, if the determinist could provide a convincing argument for the determinism of either the physical or the neural realm, he would not be required to supply a further argument for the reduction of mental realm to this realm.

Physical determinism

In the pre-quantum-mechanical era, the determinism of the physical realm seemed likely. The method of Cauchy problems, that is, obtaining the laws of physics in the mathematical form of differential equations, whose solutions are the time-evolutions of properties in terms of which the system in question is described, which, together with the initial or boundary conditions (known values of those properties at some spatio-temporal locations) have unique solutions, proved to be a very powerful tool in describing a great variety of phenomena. This paradigmatically deterministic method seemed to be the fundamental mathematical design of nature. Quantum mechanics however, as it was cast in a rigorous mathematical form by John von Neumann[162] posited a dual dynamics for the evolution of physical systems consisting of what he called

---

[161] Leibniz believed in the causal closure of physics and the irreducibility of the mental. But to maintain these two beliefs at the same time he had to claim that the mental had no role in bringing about physical events, and that it seemed as if it did because of a harmony pre-established by God between the evolution of the mental and the physical. This is of course a logical possibility, but not an interesting option for a libertarian. Neither is the idea that physical events are overdetermined by independent and independently sufficient physical and mental causes, because in that case the mental causes would be bound to cause the event that would be brought about by the physical cause anyway.

[162] Von Neumann 1955.

"process 1" and "process 2". Process 2 was the evolution of the quantum state between any two measurements. This is the solution to a Cauchy problem (with the dynamical law being the time-dependent Schrödinger equation in the nonrelativistic case). Process 1 was the indeterministic collapse or reduction of the quantum state in measurements statistically described by the Born rule, in virtue of which physical systems obtained definite observable properties. Process 1 seemed to cast doubt on determinism, as quantum mechanics was empirically superior to classical mechanics, and could explain why classical mechanics was so successful even though it was not strictly speaking right. Process 1, however, was itself problematic. Put very simply, it was hard to see how nature should know when to shift from process 2 to process 1. In solution to this problem, called the "measurement problem", different interpretations of quantum mechanics have been suggested. Some of these offer explanations to when and why process 1 should occur, some drop process 1 altogether and attempt to explain how quantum mechanics can do without it. Those who drop process 1 are deterministic interpretations. Currently the scientific community is divided on the question how quantum mechanics should be interpreted. Some of the live options are deterministic, some of them are indeterministic (as it is discussed in detail in the Appendix).

It should be noted, however, that even if a deterministic interpretation of quantum mechanics came out winning, that would not be an immediate triumph for determinism. If deterministic laws are found to describe the evolution of the matter which is not evidently in interaction with any mind, that is not in itself an evidence for determinism in the sense that interests us. Physical laws are normally tested on systems which do not involve conscious minds. Now, as long as the empirical justification for these laws come from the observation of such systems only, this evidence does not distinguish between these laws and another set of laws that differs from the previous one only in a clause that is annexed to every single law: "unless there is a conscious mind to interfere". The only way to rule out the possibility that the laws of physics are to be understood with this clause annexed to them is to observe the evolution of the matter of the brain when the mind does interfere, e.g. when decisions are being made. If it was found out that the ultimate laws of physics are deterministic and that these laws describe the evolution of brain matter when the mind associated with that brain is taking decisions,

that would be a proof of determinism. As far as our present knowledge goes, there is no proof that physics at the fundamental level would be deterministic, let alone that the fundamental physical description of brains in action would be deterministic.

Neural determinism

Some philosophers claim, however, that the evolution of brains in action is known to be deterministic, not at the fundamental physical level, but at the "neuroscientific" level. Ted Honderich claims that the brain was found to be a deterministic neural automaton:

> The first things to consider is neurons…. Our mental lives are bound up with these most important elements of our brains and central nervous systems. Each of them is a cell into which go roots or dendrites…. Out of each goes a trunk or axon…. The roots are for input to the main body of the cell, and the trunk is for output. At the end of the trunk is a synapse or connection with other items, usually roots of other neurons. The input and output are electrochemical in nature. To begin with input to a root, chemical substances called neurotransmitters are released or secreted across a synapse, and this contribute to whether the neuron gets active or not. Some chemical inputs promote activity and some inhibit it. The activity is electrical and well understood. It consists in the passage of electrical impulses to the trunk of the neuron. These impulses occur in patterns, and result at the end of the trunk in the release of neurotransmitters across synapses to other neurons. A general truth about these building blocks of the brain and the nervous system is that their operation is indubitably taken to be causal [deterministic] by just about all working neuroscientists. No question can arise about that.[163]

If this was true, it would make quantum mechanical indeterminacy irrelevant for the question we are concerned with. If physics a level below was indeterministic, that indeterminism would be confined

[163] Honderich 2002, pp. 65-66.

then to a random selection from the multiple possible physical realizations of deterministically evolving neural states. As Daniel Dennett put it, quantum indeterminacies do not result in macroscopic indeterminacies about human behaviour unless "natural Geiger counters", i.e. amplifiers of quantum mechanical effects are involved in the workings of our brains. Otherwise, our brains being large and hot systems, with many degrees of freedom, it is very likely that quantum indeterminacies cancel out on the macroscopic level.[164] If what Honderich says is true, then it proves that there are no natural Geiger counters in our brains.

Neither does arise the problem that arose with respect to physical determinism, i.e. that we have a deterministic theory that is empirically corroborated by data obtained only from the study of systems that can be thought to be isolated from the interference of conscious minds, for trivial reasons.

However, neural determinism is not nearly as uncontroversial as Honderich claims. Other philosophers claim with equal assuredness in their tone that virtually all working neuroscientists agree that there are natural Geiger counters in our brains and they have an important role in the evolution of neural states. Henry Stapp writes that

> Quantum mechanics deals with the observed behaviour of macroscopic systems whenever those behaviours depend sensitively upon the activities of atomic-level entities. Brains are such systems. Their behaviours depend strongly upon the effects of, for example, the ions that flow into nervous terminals [synapses]. Computations show that the quantum uncertainties in the ion-induced release of neurotransmitter molecules at the nerve terminals are large (Stapp 1993, pp. 133, 152). These uncertainties propagate in principle up to the macroscopic level. Thus quantum theory must be used in principle in the treatment of the physical behaviour of the brain, in spite of its size.[165]

John Eccles defended an interactive dualist picture by hypothesizing that the self controls its brain by what Stapp called

---

[164] Dennett 1984a.

[165] Stapp 2007, p. 300.

"biasing the quantum statistical rules" applicable to the quantum processes at the synapses mentioned by Stapp.[166]

Various philosophers objected against this supposition.

Honderich claimed that Eccles's theory was a hidden variable interpretation of quantum mechanics, in which the self or originator, irreducible to the physical events going on in the brain, plays the role of the hidden variable filling in the gaps in the quantum mechanical explanation for the flow of physical events. Honderich thinks that the indeterminist-interactionist like Eccles has, on the one hand, to embrace the completeness of quantum mechanics, thereby denying the possibility of hidden variables, in order to be a physical indeterminist, and then, on the other, adopt a hidden variables theory, to be in the position to deny that quantum mechanical events at the synapses are chance events.[167]

Stapp claimed that this hypothesis "upset the logical coherence of the whole scheme", because it contradicted the Born rule.[168] David Papineau discarded this suggestion on the same ground, and added that the causal closure of physics could be vindicated even if quantum mechanics was objectively indeterministic, because interference from outside would not be possible without spoiling the Born rule.[169]

Both arguments are mistaken.

Honderich simply misunderstands the sense in which quantum mechanics is taken to be complete on the interpretations that take it to be complete, as we will see it shortly analysing the Stapp-Papineau argument. It should also be clear that if, for example, the experimenter in either of the wings of an EPR-Bell experiment has a non-physical free will which interferes in the physical processes going on in the neurons of his brain to set the Stern-Gerlach magnet to detect the spin component he wants to detect is not the kind of hidden parameter the existence of which was excluded by Bell's analysis of this experimental situation.[170]

What Stapp and Papineau say has an air of obviousness. If free will is to operate in the room created by the quantum mechanical indeterminacy at the synapses, then it should mean that it brings about one of the physical states there that were quantum-

---

166 Eccles 1990, 1994.

167 Honderich 2002, p. 76.

168 Stapp 2007, p. 310.

169 Papineau 2001, footnote 2.

170 Bell 1964, 1966. For more on the EPR-Bell experiment and its relevance to hidden variable theories please refer to the Appendix.

mechanically possible, and thereby biases, that is to say contradicts, the Born rule. The Born rule, however, is about the probabilities of possible physical outcomes, and bringing one of them about by interfering from outside does not contradict it on either the propensity interpretation[171], or the frequency interpretation[172] of probability.

On the propensity account probability is a dispositional property, an intrinsic tendency of the physical system in question to produce one or another outcome. (This idea would require causation to be a deeper fact than what is captured by the nomological account.) The concept is applicable to a single trial, or even when no trial is actually made. It is an objective feature of physical reality, whether or not we test it with experiments.

On the frequency account, on the other hand, frequencies are not only empirical *evidences* of probabilities, they *constitute* probabilities. More precisely, probabilities are defined as relative frequencies in long, ideally infinite, series of repeated trials. Probabilities are, therefore, properties of mass phenomena, or long series of repetitions, and not intrinsic properties of physical systems.

If the probabilities are propensities, then they are the intrinsic physical properties of the respective physical system. If the choice the free agent makes is irreducibly mental, then it was not an intrinsic property of the respective physical system that what the agent's choice brought about in it would happen to it, for what the agent brought about in it was counterfactually dependent on something genuinely non-physical, i.e. his choice. What happened to the physical system because of the agent's self working as an originator does not speak to the question what intrinsic tendencies where there in the physical system before he interfered. His interference cannot disprove, contradict or bias any natural law about the probabilities of possible outcomes.

If probabilities are relative frequencies in a long series of trials, then singular interferences will not affect them. The frequency account can work only if the series of trials in which relative frequencies are counted are *really* long, if not infinite[173]. The change in relative frequencies brought about by any singular interference tends to zero as the length of the series of trials tends to infinity. It is true

---

171 Popper 1959.
172 Von Mises 1931.
173 Hajek 1996.

that regular interferences can affect relative frequencies and thereby probabilities. But only if not just the event of interfering obtains regularly, but there is also a regularity in what the result of the interference is. Otherwise the effect of one interference on the relative frequencies could be compensated by the effects of other interferences. There is a room for interfering regularly, a very large number of times, without compromising the probabilistic predictions of quantum mechanics, as long as there is no regularity in the interference, or if there is an even much greater number of cases when we do not interfere.

Henry Stapp, being himself a dualist interactionist, views quantum mechanics not as a theory of the evolution of the physical state of the world that allows for interference by the irreducibly mental, but as being itself a theory of the interaction of mind and matter. His interpretation of quantum mechanics is a variant of the solution to the measurement problem that became dominant first in the course of the development of quantum mechanics, emerging most prominently form the work of Bohr, Heisenberg, von Neumann, and Wigner. A summary of this interpretation is to be found in the Appendix. On this view process 2, the evolution of the quantum state described by the formalism of quantum mechanics is uninterpretable without an explicit reference to probing situations and to a conscious observer who decides which probing question is to be posed, and induces process 1 by posing that probing question physically. Posing a probing question physically is placing a measuring apparatus with which the measured system interacts. There is a conceptual duality present in the description quantum mechanics gives to the combined system consisting of the measured system, the measuring apparatus and the experimenter. The latter two are described in classical terms, while the first is described in the formal mathematical language of quantum mechanics that is uninterpretable in classical concepts apart from the context of the measurement situation. The separation of these two parts of the total picture is called the von Neumann cut (or Heisenberg cut). As it was often emphasized by Bohr,[174] this conceptual duality was impossible to overcome. Yet, it seemed arbitrary to cut nature into two halves and describe the two halves in two fundamentally different ways. Why not include the physical measuring apparatus in the half that is to be described quantum

[174] Cf. Bohr 1958.

mechanically? Von Neumann proved[175], thereby removing an alarming ambiguity from the formalism, that, although the existence of the cut was essential to the formalism, its location was indeed arbitrary. It could be anywhere, without there being any difference in the predictions of the theory. This result is standardly called "the moveability of the von Neumann cut". Stapp suggests that the cut should be moved "all the way up":

> [T]his cut could be pushed all the way up so that the entire physically describable Universe, including the bodies and brains of the agents, are described quantum mechanically. This placement of the cut does not eliminate the need for Process 1. It merely places the physical aspect of the Process 1 psychophysical event in the brain of the conscious agent, while placing the conscious choice of which probing question to pose in his stream of consciousness. That is, the conscious act of choosing the probing question is represented as a psychologically described event in the agent's mind, which is called by von Neumann (1955, p. 421) the "abstract ego". The choice is physically and functionally implemented in his brain. The psychologically described and physically described actions are the two aspects of a single psychophysical event, whose physically described aspect intervenes in the orderly Process 2 evolution in a mathematically well defined way.[176]

Without this work done by the irreducible mind the dynamics of the evolution of the physical world would be fundamentally incomplete. The gaps in it are filled in by free choices made by conscious agents. With it, quantum mechanics is a dynamically complete, though indeterministic, theory of the evolution of the physical world interacting with our consciousness.

Stapp refers to some psychological phenomena that his theory explains better than rival theories, including the effort of attention, which he explains as a case of the quantum Zeno effect.[177]

From the perspective of freedom Stapp's interactionism is less satisfying than Eccles's, given that on Stapp's account there is no

---

[175] Von Neumann, 1955.
[176] Stapp 2007, pp. 304-5.
[177] Ibid, p. 307.

controlling which physically possible state should obtain in process 1. The mind controls only the range of possibilities, that is, chooses the basis (the set of eigenstates) in terms of which the pre-process 1 state should be expounded as a superposition state, by choosing a probe question, but does not control the outcome (the eigenstate to which the superposition state actually reduces). Stapp's account is nevertheless as indeterministic as Eccles's. (And shows how interactive dualism could be true, even if Papineau was right claiming that any tempering with outcomes was ruled out by the Born rule).

I see no reason why the mind could not do the work that Eccles attributes to it if it does the work that Stapp attributes to it. We have seen that, contrary to Stapp's claim, it would not contradict the Born rule. It should be noted, however, that whereas Stapp's picture assumes a certain interpretation of quantum mechanics, Eccles's suggestion could be combined with any interpretation that is objectively indeterministic, for example with the GRWP interpretation, that attributes the collapse of the superposition state to entirely physical stochastic processes, without any reference to a conscious observer. (The GRWP interpretation modifies the Schrödinger equation in order to have a single dynamical evolution involving stochastic collapses. See the Appendix.)

I conclude that as far as physics and neuroscience presently go determinism is a possibility but not a near scientific fact.

7 Can We Have Alternatives Anyway? – B. On the Flow of Time

Logical fatalism

Aristotle famously discusses an argument from the law of the excluded middle to fatalism in the ninth chapter of *De interpretatione* (and also in *Categories*), and claims that the principle cannot (always) be applied to propositions about the future, there are future contingencies, and propositions describing them have no definite truth-value, and so the argument is mistaken. I agree, however, with Peter van Inwagen[178] who finds this claim very counterintuitive. I would challenge the argument at the point that the truth of a proposition yesterday is not a *past fact* in quite the sense in which ontological fixity would apply to it, and therefore it can be changed today, if the fact that it describes will obtain only tomorrow. I agree with van Inwagen that it is not *past* in quite the same sense (the truth of a proposition is timeless rather than past, present or future in the sense that facts that *become* are past, present and future), and I would add that also they are not *facts* in the sense in which facts that become real at some time are facts.

Let me expand on it a bit.

In Aristotle's classic example either of the following two propositions, that *there will be a sea battle tomorrow*, or that *there won't be a sea battle tomorrow*, is true. Whichever of these two is true, when we say it is true, we mean it is true *now*. If it is true now, then nothing, not even the deliberations and decisions of the two opposing admirals, or of anybody else, about whether there should be a sea battle tomorrow, can make it false between now and tomorrow. (The present time is not particularly relevant: the proposition in question has always been true. So no one has *ever* had a choice about whether there will be a sea battle tomorrow.) An analogous argument could be given in respect of any proposition about the future. The upshot is that no one can do anything about any future event, one's own actions included, because either the proposition that it will happen, or

[178] 1983, Chapter II.

the proposition that it won't, is already (has always been) true. So the future is not open. So there is no libertarian freedom.

I think this argument is confused. Propositions are meant to express facts. If we think there is no fact yet about whether there will be a sea battle tomorrow, then it should be considered as part of the data that needs to be taken into account theorising about propositions that the verdict that the proposition that *there will be a sea battle tomorrow* is either true or false now should not entail that no one can do anything about whether there will be a sea battle tomorrow between now and tomorrow. As long as the former seems to entail the latter, there is something wrong with how we conceive of the truth or falsity of propositions about the future.

But of course it is not a proper argument against logical fatalism unless I identify the mistake in the logical fatalist argument. So I will try to identify it.

Let *p* denote the proposition that event *e* will happen at a future time *t*. It seems plain that either *p* or non-*p* is true. We don't know which. We know that either *p* or non-*p* is true because we know that at the latest at *t* there will be a fact that will make one of them true and the other false. It is not possible that by *t* there won't be such a fact. It is guaranteed by what we mean by an event happening at a time.

So far we are in accord with the fatalist. So far we relied only on a very minimal theory of propositions: that they grasp, and are made true by, facts. Apart from this, we relied only on simple ideas about events and time. It is hard to believe that we are mistaken about any of these things. But maybe these things only guarantee that either *p* or non-*p* will be true at *t*. Maybe when he claims that either *p* or non-*p* is true *now*, the fatalist tacitly invokes some theory of propositions that goes beyond the idea that they express, and are made true, by facts.

That theory would be that if a proposition is true, then it is true timelessly. Could it be the source of the mistake?

Well, the alternative to this theory would be that a proposition has no truth-value (or has a third truth-value that we might call 'indeterminacy') until some facts or states of affairs make it true or false. On this second theory, the proposition, for example, that there will be a sea battle tomorrow has no truth-value now, if whether there will be one depends on some libertarian free choices of the two admirals that they haven't yet made. If their choices will be

necessitated by psychological and circumstantial facts that already exist, then, of course, the proposition has a truth-value.

Is the choice between these two theories the point to attack the logical fatalist argument (as Aristotle himself suggested)? I don't think so. The second of these two theories is very unattractive intuitively. Suppose *e* does happen at *t*. Suppose we are at *t* now. Suppose someone asserted *p*, the proposition that *e* would happen at *t*, at an earlier time t_0. In retrospect we would say that he was right. Our judgement at *t* about his proposition *p* he asserted at t_0 would be that it had a truth-value, namely, it was true. It would be very unnatural to say that *p* had no truth-value at t_0, now that we know that everything happened the way it was predicted by *p*.

An analogous reasoning can convince us that *p* had a truth-value at t_0 also if *e* fails to happen at *t*. The field of options is exhausted by these two. So at *t* we would very probably say that *p* had a truth-value at t_0 anyway.

Suppose we are at t_0 now and *p* is a proposition about a future time *t*. Why should we think that it has no truth-value now, given that we know that looking back from *t* we will think that it had one?

I think common sense dictates that we should agree with the fatalist that propositions are true or false timelessly. The mistake must lie somewhere else.

But it is not very difficult to locate. Suppose that *p* is true now. Does it follow that there is nothing anybody could do about *e*? Is *e* inevitable? Does it follow from *p*'s present truth that *e* cannot be the consequence of a libertarian free choice of an agent to be made between now and *t*? The answers to these questions are definitely in the negative.

This is easy to see. Suppose that *e* is a consequence of a libertarian free choice of an agent to be made between now and *t*. Suppose *e* is contingent relative to what is metaphysically real presently. Suppose time really flows, future events differ ontologically from present and past events in that their identity is not fixed yet, provided that they are metaphysically contingent relative to the present. All these suppositions are perfectly compatible with *p*'s being true now.

If these suppositions are true, then asserting *p* now is just a guess. A guess can be true if lucky. No one knows that *p* is true, not even an ideal knower, God or a Laplacean Deamon. Of course, whether *e* is contingent relative to what is presently real is not an epistemological matter, it is a metaphysical one. From the fact that *p* is not known by

anyone it does not follow that there is no metaphysical fact corresponding to *p*'s truth. There will be one. But it is *not real, it doesn't exist yet*. That is what I wanted to highlight by appealing to the ideal knower's ignorance of the truth of *p*. The metaphysical fact that makes *p* true lies in the womb of the future. This fact is that *e* will happen, although it could fail to happen, given everything that is *already real*. That is sufficient for *p*'s truth now, nothing more is necessary.

Consequently, nothing more can be deduced from *p*'s present truth.

The fatalist would of course protest against our supposition that *e* can be the result of a libertarian free choice to be made some time between now and *t*, even though *p* is already true. The mistake he commits can be unveiled if we consider how he would argue against this supposition. *It is the case now*, he would say, that *p* is true. If anyone could have a libertarian free choice about *e* at a later time between now and *t*, that would mean that this person has a libertarian free choice about what *was the case at a time earlier than the time of his choice*. And that is impossible.[179]

If the fatalist thinks it is impossible, it is because he believes that past facts cannot be undone. If a fact is classified as 'past' temporally, it means it is also classified as 'unchangeable' ontologically. This belief is at the heart of his argument.

I share this belief with the fatalist, yet, I think he is mistaken.

When he was arguing (or we were arguing on both his and our own behalf) against the theory that propositions have no truth-value before the events they describe actually happen (or before causally necessitating conditions for it occur), he argued in effect that the fact that makes a proposition true can be in the future, it doesn't need to be *already real* for the proposition to be *already true*. And now he seems to be saying that the proposition's being *already true* entails that the event the proposition describes, that is, the fact that will make the proposition true, is *already real*, and so cannot be undone.

Surely, he can't have it both ways.

But at this point the fatalist might protest that I am misrepresenting his position. He is not committed to the view that future events differ ontologically from past and present events. He is

---

[179] The credit for first presenting the fatalist argument as an argument from the unchangeability of the past is traditionally given to Diodorus Cronus. (See Epictetus, *Dissertationes* II 19, 1-5 in Döring 1972.)

comfortable with the position that all events share the same ontological status, they are all equally real and, therefore, unchangeable. So he is not committed to the view that propositions about the future are made true by something *not yet real*. So he is not contradicting himself the way I am accusing him.

There are two problems with this line of defence from the fatalist's part, and both are destructive to his position.

The first is that if he believes that future events are as real as present and past ones, then he believes that existence is timeless. But if existence is timeless, agents are timeless, too. If this is so, then, of course, the temporal modifier in the question whether an admiral has a choice *between now and tomorrow* about whether there should be a sea battle tomorrow refers only to a position in phenomenal time, it does not refer to an ontological situation, i.e. the situation that *it is already the case* that the proposition that there will be a sea battle tomorrow is true. If existence is timeless, then temporal qualifications do not apply to statements describing what is real. A timeless admiral can have a libertarian free choice about the battle *timelessly*. Maybe the event of the sea battle tomorrow is just as real as my headache yesterday, yet it may be the result of libertarian free choices made by naval commanders. The fact that, in phenomenal time, the relevant choices are made between now and tomorrow, and the fact that the sea battle is as real as any past event do not contradict each other. If existence is timeless, then temporal ordering in phenomenal time bares no ontological consequences. It may be the case that the timeless choices of the admirals are to unfold in phenomenal time between now and tomorrow. They have a choice about whether the battle should happen, a timeless one, but in the tentative language of phenomenal time it can be said that they have a choice about it between now and tomorrow. (More will be said about this matter a little later.) This is contradictory to fatalism. So I think the fatalist had better keep a his position apart from the position that existence is timeless.

The second problem is that either the fatalist has to rely on some argument to the effect that propositions are true timelessly even if future events are ontologically different from present and past ones, or, alternatively, he has to provide an argument, not relying on the thesis of the timeless truth or falsity of propositions, to the effect that future, present and past events share the same ontological status. Otherwise his argument for fatalism won't get off the ground. If he chooses the first option, then he will find himself committed to the

view that propositions can be made true by facts which are *not yet real*, so when he will argue that the power to make true propositions about the future false would require the power to undo facts which are *already real*, he will find himself in the contradiction with which I am accusing him. If he chooses the second option, then he makes his position vulnerable to the objection that timeless existence may be hospitable to libertarian freedom, as it was presented in the previous paragraph.

But maybe the fatalist can object to the charge of incoherence another way. I said that when he claims that propositions are true timelessly, he commits himself to the view that propositions can be made true by facts, states of affairs, which are *not yet real*, and when he claims that having a libertarian free choice about a future state of affairs would require the power to undo facts that are *already real* because it would require the power to make an already true proposition false, he is contradicting this first commitment of his. Now he may protest that when I said that he was contradicting himself I did not realize that he was not treating the same facts or states of affairs as *not yet real* in the first commitment he made, and *already real* in the second. They are different states of affairs, he might say. The one that is *not yet real* is *e*'s happening. The one that is *already real* is not the fact that makes *p* true (*e*'s happening), but *p*'s truth itself. So there is no contradiction in what he is saying.

But can the fact that *e* happens at *t* and the truth of the proposition that *e* happens at *t* considered as two distinct elements of metaphysical reality, of which one is capable of entering existence and thereby acquiring ontological fixity while the other is still looming in the void that is called the future?

I think this proliferation of elements of metaphysical reality (facts, states of affairs) is better to be avoided. Regarding a proposition's truth as a fact in its own right, distinct from the fact that makes the proposition true, would require treating propositions as individuals of our ontology, and then their having a predicate, namely truth, would be a state of affairs in its own right. But I think it would be quite unnatural to hold that propositions are part of the fundamental ontology of the world, it is much more sober to say that propositions are not individuals in their own right, and so their truth does not constitute an element of metaphysical reality, a state of affairs, in its own right either. The truth of *p* is not a distinct state of affairs, metaphysically speaking, over and above *e*'s happening at *t*. So the

truth of *p* now should not be treated *as a state of affairs* that obtained earlier than *t*.

In some sense *p*'s truth is of course a 'fact', and that 'fact', no doubt, has already obtained. It has already obtained when the world was born. And it will always be a 'fact'. The timelessness of this 'fact', and of all similar 'facts', i.e. the truth of propositions about metaphysical facts that obtain at some point of the history of the universe, shows that this 'fact' has nothing to do with the ontologically significant temporal classification of facts according to which a fact's being past entails its being unchangeable. That ontologically significant classification, if it applies at all (and the case when it doesn't apply at all has already been discussed), applies only to metaphysical facts that underpin 'facts', i.e. to the states of affairs that make the propositions about them true. So it is illegitimate to refer to the truth of *p* as "a fact that has already obtained and so cannot be changed".

After all, the power to make *p* false is the same thing as the power to prevent *e* from happening. If the truth of *p* was a distinct state of affairs, a fact, not just a 'fact', then probably, the power to prevent it from obtaining would require something over and above preventing *e* from happening. This extra requirement would include the power to make a difference to the past, since, if the truth of *p* would be a state of affairs in the metaphysically relevant sense, then it would be one such state of affairs that have already obtained. But it is obvious that preventing *e* from happening is sufficient for making *p* false, and preventing *e* from happening, in itself, does not require anything like making a difference to the past, the part of metaphysical reality that is already laid down, since *e* is in the future.

The truth of *p* now is compatible with someone's having the power to prevent *e* from happening. It only requires that this power be not exercised. But we are all familiar with powers which are not actually exercised yet perfectly real.[180]

---

[180] Cf. van Inwagen 1983, p. 42 ff. Van Inwagen discusses the objection (Richard Taylor's) that one cannot really be said to have a certain power if one can never exercise it. The impossibility to exercise this power is thought to be demonstrated by Taylor by the demand he thinks is obviously impossible to meet: "pick any true proposition about the future, and then so act that this proposition [is and] has always been false". I think van Inwagen satisfactorily answered this objection so I omit the answer here.

I consider the argument I am presenting in this section as a close relative of van Inwagen's argument against fatalism as he gave it in Chapter II of his 1983. His argument was based on the claim that when we say that a proposition is true *at a time*, we

Getting back to the classic example, even if it is true now that there won't be a sea battle tomorrow, it may be perfectly possible that there would be one. There could be one, it is just that, yet unknown to anyone, even to an ideal knower, there won't. It is perfectly possible that the lack of a sea battle tomorrow will be attributable to the free choices of the two opposing admirals, which they make after a whole night of anxious deliberation, and which they could make differently. They have the power to make the proposition that there won't be a sea battle tomorrow false, even if it is true now. If the proposition is true, it is true because they won't decide otherwise, and not the other way around: it is not true that they will not, or could not, decide otherwise because there is a now-true proposition that binds them. The order of explanation and the direction of dependence between facts and 'facts' is not that way.

McTaggart and Parmenides

Following a very influential article by Ellis McTaggart, it is commonly held that there are two broad ways of thinking about time. This is how McTaggart described these ways:

> Positions in time, as time appears to us *prima facie*, are distinguished in two ways. Each position is Earlier then some, and Later then some, of the other positions. And each position is either Past, Present, or Future. The distinctions of the former class are permanent, while those of the latter are not. If M is ever earlier that N, it is always earlier. But an event, which is now present, was future and will be past.

---

use the temporal modifier in a different sense from when we say that a certain state of affairs obtains at a time. He says the fatalist's argument is based on the equivocation of the temporal modifier. I agree. My argument could be easily translated into van Inwagen's language. I argued essentially that the truth of a proposition should not be considered as a state of affairs that obtaines at a certain time. We can talk of a proposition as being true at a certain time but with that we are not referring to a distinct metaphysical fact of the matter over and above the fact grasped by the proposition. If we proceed as if we did, in an argument about whether we have the power to change "what has already been the case", then we commit the fallacy of equivocation van Inwagen was talking about. I think my argument is more direct and easier to follow, but others may think otherwise.

[...]

> For the sake of brevity I shall speak of the series of positions running from the far past through the near past to the present, and then from the present to the near future and the far future as the A series. The series of the positions which runs from earlier to later I shall call the B series.[181]

McTaggart believes two important things in relation to these two series. One is that time cannot be real without being an A series, besides being a B series. The other is that the conception of the A series is incoherent. His infers from these two theses that time is unreal.

For the issue of libertarian freedom the question whether the A series is coherent may be important in its own right, whether or not it is true that time has to be an A series to be real. It is because there is no becoming in a B series which is not an A series, and *prima facie* becoming is a necessary condition for libertarian freedom. The one-place predicates of positions in the A series, Past, Present and Future, have to be real besides the two-place predicates of the B series, Earlier and Later, for becoming to be real. For becoming takes place in the present, and is conceived as a transition between an ontological status that corresponds to the predicate 'Future' and another one that corresponds to the predicate 'Past'. In an A series the future can be open in the sense that seems essential for libertarian freedom, while a B series, without being an A series, corresponds to a static universe which is a solid four-dimensional block of events of the same ontological status (though it may sound a bit technical to talk of events in such a universe, as nothing ever really *happens* in such a universe).

To say that the A series is unreal is the same thing as to say that becoming is unreal, and the same thing as to say that change is unreal. For by change we mean that something acquires a property which it hasn't had before, or looses a property which it previously had, or that something starts or ceases to exist. And this is exactly the idea that something becomes something that it was not before. And this is possible only if the temporal properties of the positions in an A

---

[181] McTaggart 1908, p. 458.

series: being future, being present, and being past, are real and correspond to the ontological states: not existent yet, existent, not existent any longer.

As of philosophers who previously held his position about the unreality of time, McTaggart refers to Spinoza, Kant, Hegel and Schopenhauer[182], but he says he supports his position with reasons none of these philosophers employed. I think both his position and the reasons he offers in support of it makes him akin to a philosopher, a founder of a much older tradition, he doesn't mention. McTaggart thinks that the A series is unreal because he thinks the very idea of the A series, and of change, and of becoming, involves a contradiction. To me it seems that his argument to this effect has very much in common with the argument Parmenides of Elea offered to the very same conclusion, and I think the two arguments stand or fall together.[183]

Central to the Eleatic tradition is that change is an illusion of the fallible senses of mortals, which nevertheless can be overcome by Reason, which reveals that whatever exists exists without beginning and end, unchanging, and thus reality is an ontologically homogeneous block. Reason reveals this analytically, by showing that change cannot be conceived coherently.

Parmenides's main argument to this conclusion is concerned with two mutually exclusive predicates that, he believes, cannot be predicated of the same subject: existence and non-existence.

We account for change as something happening to a subject. But a change means that something becomes something else. So, strictly speaking, we talk of two subjects when accounting for a change: the thing before, and the thing after the change. The first ceases to be, the second comes to be in the event of the change. Either of these two is considered, however, in our account of change the predicates of existence and non-existence will equally apply to it, and this is a problem, for these are incompatible predicates.[184] Predicating

---

182 P. 457.

183 To give a historically correct reconstrution of Parmenides's views is beyond both my competence and ambition. My aim with referring to Parmenides is only to illustrate what I believe to be the flaw in McTaggart's argument, which is considered as dubious but still live, i.e. yet to be refuted, by many even today. Nevertheless, judged by his fragments, which, unfortunately, I can access only in translation, I think the argument to follow is correctly attributed to Parmenides. Of course I may be wrong.

184 Cf. Kirk, Raven and Schofield (1983), **293**, **294**, **296**, especially **293**, 7-9 and **296**, 19-21.

mutually exclusive predicates of the same subject is a contradiction. So change cannot be conceived coherently.

At this point someone might come forward saying that Parmenides cannot really mean this. For it is evident that the two mutually exclusive predicates are not predicated of the subject (either subject) *at the same time*. The thing that comes to be in the change *was* nonexistent before the change, and *is* existent now that the change has already taken place. The thing that ceases to be in the change, the other way around. So it is *never* the case that the same thing is predicated to be both existent and nonexistent.

But Parmenides, I think, would not find this objection impossible to deal with.

First of all, he would answer that it is enough trouble for the advocate of becoming that on his account of change (there is a time when) noexistence is predicated of something. Parmenides seems to think that predicating non-existence of any subject whatsoever is a contradiction in itself, even if existence is not predicated of the same subject.[185]

Now I think this is an absurd position, which entails that we cannot question the reality of chimeras or unicorns, and which invites a quite trivial version of the ontological argument for the existence of God, for example, making the whole effort invested in the argument by philosophers from Anselm to Plantinga completely redundant. But we can set this aside for the moment, because Parmenides might come up with another answer.

He might say that his opponent is trying to remove the contradiction he pointed out in the doctrine of becoming with an appeal to different *times* at which non-existence, on the one hand, and existence, on the other, can be truly predicated of the same thing. But Parmenides might protest that different times exist only if becoming is real. Indeed, many philosophers believe that there is no time without change, even if the way McTaggart puts it, "it would be universally admitted that time involves change", is overstating a bit the consensus about this matter. But it is conceivable that Parmenides believed that time is unreal, and he believed it because he was convinced that change and becoming are unreal, and he believed that time involves change. If so, then he might answer to his opponent that this appeal to different times is illegitimate, as long as the very

[185] See KRS 291.

existence of time is in doubt, because of an apparent contradiction that has been pointed out in the idea of becoming. Becoming is essential to time, but the apparent contradiction in the idea of becoming cannot be removed if not by an appeal to time, as if the idea of time were undoubtedly coherent. The thought of Parmenides's opponent moves in a circle, and this circle is a vicious one. Or so Parmenides might claim.

Well, first of all, for the charge of circularity to stand against Parmenides's opponent, it must be true that "time involves change". If time is not dependent on change, there is surely no circularity in the way Parmenides's opponent removes the contradiction Parmenides thought to have found in the notion of becoming. I think Parmenides's opponent can be defended against the charge of circularity even if it is accepted that there is no time without change, i.e. without becoming, so this question is not absolutely essential to our discussion, but it is a fascinating one, deserving attention in its own right.

There is a metaphysical and a conceptual version of the claim that time involves change. The metaphysical version of the thesis asserts that in a universe without change there is no time. The conceptual version of the thesis asserts that we cannot conceive of time without relying on the notion of change, if the latter concept proves to involve a contradiction, so does the former. If the charge of circularity stands against Parmenides's opponent, it is not sensitive to whether we understand the thesis in the metaphysical or in the conceptual sense. Parmenides thinks his opponent commits the fallacy of circularity, because if time is metaphysically or conceptually dependent on change, and change is conceptually dependent on becoming, and the concept of becoming involves an apparent contradiction which Parmenides's opponent tries to remove with an appeal to time, then this move of his seems to presuppose the conclusion he desires to reach, i.e. that the concept of becoming is free from the contradiction that appeared to Parmenides.

Although it might seem quite uncontroversial that time is dependent on change, since whenever we measure time we compare the temporal distance between two events to a rate of a change, it has been debated ever since the time of Plato and Aristotle. The debate over the dependence of time on change has been entangled with the issue whether time (like space) is a container in which events take

place, a container which is logically and metaphysically prior to any event (this position is often called the "substantival" view of time), or time (like space) is a system of relations between events which does not exist without there being events bearing temporal (as well as spatial) relations to each other (this is the "relational" view). By the entanglement of the two issues I mean that a relationalist can be expected to hold that time is conceptually dependent on change, because he might think that without change, without anything happening, there can hardly be any temporal relations between events, on which the concept of time could be based. A substantivalist, on the other hand, is likely to deny the dependence of time on change, either metaphysical or conceptual, since he holds that time has to exist already for any event to be possible. Plato was of course a substantivalist, and Aristotle was a relationalist. On the advent of modern physics, the two inventors of its revolutionary mathematical language, in which infinitesimally small changes taking place in infinitesimally small periods of time, Newton and Leibniz, disagreed famously on the same question, Newton fiercely defending the "substantival", Leibniz the "relational" view of space and time. In twentieth century physics the theoretical challenge posed by the empirical invariance of the speed of light[186] gave rise, again, to parting interpretations, Lorentzian mechanics being on the substantivalist side, and Einstein's special theory of relativity being on the relationalist side.

It should be noted, though, that Einstein's view is an example for that the entanglement of the two issues (substantivalism vs. relationalism, on the one hand, time without change vs. no time without change, on the other), should be handled with caution, and the two issues should not be conflated. The special theory of relativity is definitely relationalist about time but on the dominant view it lacks change in the McTaggartian sense, although it is not accepted unanimously.[187]

Although the theory of relativity is a very successful discipline of modern physics enjoying nearly unanimous acceptance in the scientific community, substantivalism about time (and space) is not completely dead. More will be said about this matter a little later.

---

[186] Strictly speaking there is empirical evidence onty for the invariance of the *two-way* speed of life, as we will see later.

[187] It is also true though that on the dominant view the time of STR *does not flow*. On the minority view it does flow in a local way.

In the relatively recent philosophical literature on the topic, Sidney Shoemaker gave a new twist to the discussion by describing a conceivable world in which there are times when no change is taking place, yet time flows.[188]

Shoemaker's world is divided into three thirds by glass walls. Let these thirds be called Region A, Region B, and Region C. Each region is inhabited. People belonging to different regions do not mix but can communicate with each other. Each region is observable from the other two. After the first two years have passed without any apparent irregularity in Shoemaker's world, the inhabitants of Region A observe a discontinuity in the life of the other two regions. At midnight, 31 December of the second year, the picture they see over the glass walls changes suddenly. There is a leap. What they see in the first moment of the new year is not continuous with what they saw in the last moment of the old year. Their own region however they perceive to evolve smoothly. They inquire about what might have happened from their neighbours. The inhabitants of both Region B and Region C report that there was no discontinuity in their lives, but they observed that life in Region A froze for a whole year in the third year, and now it is the fourth year of the world. In that year literally nothing happened in Region A as if someone pushed the pause button on a remote control unit to press play again only after the third year has passed. So the third year of the lives of Region B and Region C people passed unobserved by Region A people. In their perception the last moment of year two was immediately followed by the first moment of year four. This is why they perceived a discontinuity in regions B and C.

Even if Region A people have some doubt about this explanation, it clears up at the end of the fourth year, (the third in their perception), when they see that life in Region B freezes for a whole year. The same thing happens in Region C a year later. As years pass, and the people of all three regions see the other two regions freeze again and again on a regular basis, and observe discontinuities in the other two regions of which the most credible explanation is that they themselves freeze, as well, regularly, they all come to the conviction that Region A freezes in every third, Region B in every fourth, and Region C in every fifth year, for a year. This cosmological theory explains all the strange phenomena they are confronted with.

---

[188] 1969. I present a slightly simplified version of Shoemaker's thought experiment in the hope that it does not compromise the philosophical lesson to be drawn from it.

Now do these freezes constitute cases of time without change? Hardly so, because when one region freezes, its time is the same as that of the other two, and the time of the other two regions is time with change. But consider what happens in every sixtieth year. What the people of all three regions observe is that the regular freezes they have so far observed in the neighbouring regions fail to occur. They might find it surprising first that a phenomenon that so far has recurred with great consistency now fails to do so. But upon little reflection they might come to the idea that it probably did not fail to occur, after all. They failed to observe it, because—given the recurrence of freezes they have observed so far—in every sixtieth year the freezes in the three regions are expected to occur simultaneously. Phenomenally the concurrence of freezes presents itself exactly as if no freeze had occurred in any of the regions. Since everybody is frozen while everybody else is frozen, no one observes either a freeze or a leap in the life of the other. The inhabitants of Shoemaker's world have all the reason to believe that, in every sixtieth year, a whole year passes in their world without any change taking place in it, while all three regions are frozen.

Now is it a case that disproves the principle that "time involves change"? In some sense it is. But I am not sure if it is the sense which is relevant to our problem. The belief of the inhabitants of Shoemaker's world in the year that has passed in their world without any change occurring in it is based on a belief in the reality of time they developed the same way as we, the inhabitants of this world, do. It is grounded in a belief in the reality of change. The time of the sixtieth year is an extrapolation of the time of the first fifty-nine, when there have always been change in Shoemaker's world, at least in one of its regions.

But is it true, metaphysically speaking, that that one year long period, when the whole world stands still, could not be there, could not possibly exist, if the first fifty-nine years had not been there? Yes, it is true that the awareness of Shoemaker's people of that year is dependent on there being change in other years. But is it true that the mere existence of the year is dependent on that, as well? It is somehow strange to assume that a world can stand still only if it moved before. If it is not true, then isn't it conceivable that there is a world whose time consists of periods like the sixtieth year of Shoemaker's world, glued one after the other, say, in an infinite sequence?

Supposing that a completely frozen world, lasting from the infinite past to the infinite future, is possible, does it prove Parmenides wrong? I think we should understand the conclusion of Shoemaker's story—if it is the right conclusion to draw from it—with the clause: "provided that our concept 'time' makes sense at all". If this story proves that time can pass in a world that is motionless from infinity to infinity, it does so only on the condition that time in normal worlds is a coherent idea referring to a real feature of these worlds, not just an illusion based on a confusion. If time is a container of events (events understood in the sense that involves becoming), then the time of a motionless world is an empty container. As such, *it has the potential to contain events*, it is just that actually it doesn't contain any. Now if the idea of an event, which involves the idea of becoming, is a confusion, then the idea of containing events, either potentially or actually, is a confusion too, even if thinking of a container that doesn't actually contain any adds no further confusion to the already existing one.

So I think we may conclude that Shoemaker's world does not disprove the principle that "time involves change" in the conceptual sense, the sense that is relevant to our discussion.

Since I am unaware of any other suggestions that would provide any hope to definitely disprove the principle that "time involves change", in the required sense, I abandon this line of defence on behalf of Parmenides's opponent against the charge of circularity.

For the purposes of the discussion to follow I propose that we proceed as if the principle was undoubtedly correct. With this I give an advantage to Parmenides, whom I want to prove wrong. I think his opponent can be cleared from the charge of circularity, even if this principle holds.

Circularity, in the philosophically bad sense, is the case when someone is trying to convince us of something with an argument, and that something, the conclusion of the argument, needs to be included among the premises for the argument to go through. Does Parmenides's opponent really commit this mistake? I don't think so. I think it is Parmenides, who argues circularly. What his opponent is doing is pointing out the circularity in Parmenides's argument.

Parmenides's opponent is not trying to explain what time is, or what becoming is. If it is true that time is conceptually dependent on change, and so on becoming, then in his account of time he would

inevitably mention becoming, or some synonymous notion. If it is true, what Parmenides seems to suggest, that an account of becoming inevitably involves a reference to time, or temporal determinations, then—provided that neither time, nor becoming is considered as ultimate—the explanation of what time and becoming are will be circular. But that does not necessarily constitute a philosophical problem. Time and becoming may be twin-concepts, mutually explaining each other in a circular way, and the pair of them taken together being ultimate, admitting of no explanation in terms of other concepts in a noncircular way. There is nothing wrong with that philosophically. I think the appeal to time by Parmenides's opponent, when he is trying to explain why it is that the concept of becoming does not involve the contradiction of predicating incompatible determinations (non-existence and existence) of the same thing, is ill-represented as a circular, and so failed, attempt to account for these two notions.

Nor is Parmenides's opponent arguing that the concept of becoming is coherent. He is not trying to prove that. If he tried to prove that by an appeal to the notion of time, without reflecting on time's dependence on becoming, then his argument would be circular. But he is not attempting a proof of this manner.

All he is attempting is to show that *there is no contradiction in the notion of becoming unless we assume that time is unreal.*

We are operating now under the assumption that if becoming is unreal, then time is unreal, too. Yet, I think, it is misleading and tendentious if Parmenides describes his opponent's move as "trying to remove the contradiction by assuming that time is real, (which is equivalent to assuming that there is no contradiction in the notion of becoming, because if there is one, then time cannot be real)".

For, unless we assume the unreality of time, there is no contradiction in the notion of becoming to start with. The appeal to time by Parmenides's opponent is not for "removing a contradiction", rather, it is pointing out that we need to assume that our perceptual consciousness fails us, and there is no time really, to have a contradiction in the first place. That is a quite unnatural assumption, one that needs to be supported with good argument. Now it seems that all Parmenides has to offer in support of the thesis of the unreality of time is the alleged contradiction he thinks he pointed out in the notion of becoming, in combination with the thesis that time cannot be real without becoming being real. But this

argument is, of course, circular. It needs to have the intended conclusion, time's unreality, among its premises, otherwise there is no contradiction in the notion of becoming. If the claim that time is unreal would be supported by independent reason, then Parmenides would have a noncircular argument for the unreality of becoming. But it seems that we have no reason to think that time is unreal, unless becoming is unreal. At least, no such reason has been put forward by Parmenides to think that time is unreal, independent of the question of the reality of becoming. So as an attempted proof of the unreality of becoming, the argument presupposes its conclusion, so it is circular in a bad sense. Parmenides's opponent appeals to time only to make this circularity visible.

Now, Parmenides does prove something interesting about the idea of becoming, nonetheless. He proves that the idea of becoming is an unusual idea. It is unusual, because *it needs to have a real referent to be coherent*. Most concepts are not like that. The concept of the unicorn may be perfectly coherent without there ever being a real unicorn. But from this unusuality of the idea of becoming, however, it doesn't follow that it is incoherent. If there is becoming, then there is time, and then the idea of becoming involves no contradiction whatsoever.

Although I believe that the position I have attributed to Parmenides in this imaginary exchange between him and his opponent is not alien to what Parmenides might have really thought, (as it was indicated earlier in footnote two or three pages above) it is not my intention to claim anything about the historical Parmenides. My sole purpose with this story of my imaginary Parmenides is to sink McTaggart's argument. My Parmenides is almost McTaggart. McTaggart's argument is a little more complicated, so it is a little less obvious to see the mistake in it, but it has the same structure, and it rests on the same mistake.

This is how McTaggart argues for the unreality of the A series.

> The terms of the A series are characteristics of events. We say of events that these are either past, present, or future.
>
> [...]
>
> Past, present, and future are incompatible determinations. Every event must be one or the other, but no event can be

> more than one. This is essential to the meaning of the terms. And, if it were not so, the A series would be insufficient to give us, in combination with the C series, the result of time. For time, as we have seen, involves change, and the only change we can get is from future, to present, and from present to past.

So far the argument runs parallel with the Parmenidean one. We are told that the idea of the A series requires the predication of incompatible determinations of the same subject, as did the idea of becoming in the previous case.

Now this alleged problem invites the same explanation as did Parmenides's worry that the idea of becoming involves the predication of both non-existence and existence of the same thing.

McTaggart continues:

> It may seem that this can easily be explained. [...] It is never true, the answer will run, that [an event] M *is* present, past and future. It *is* present, *will be* past, and *has been* future. Or it *is* past, and *has been* future and present, or again *is* future and *will be* present and past. The characteristics are only incompatible when they are simultaneous, and there is no contradiction to this in the fact that each term has all of them successively.

Against this explanation McTaggart raises the charge of circularity as did Parmenides above.

> But this explanation involves a vicious circle. For it assumes the existence of time in order to account for the way in which moments are past, present and future. Time then must be pre-supposed to account for the A series. But we have already seen that the A series has to be assumed in order to account for time. Accordingly the A series has to be pre-supposed in order to account for the A series. And this is clearly a vicious circle.
>
> [...]

> We have come then to the conclusion that the application of the A series to reality involves a contradiction, and that consequently the A series cannot be true of reality.

McTaggart readily assumes the objection the analogue of which has been put forward against Parmenides, and answers it.

> We must consider a possible objection. Our ground for rejecting time, it may be said, is that time cannot be explained without assuming time. But may this not prove—not that time is invalid, but rather that time is ultimate? It is impossible to explain, for example, goodness or truth unless bringing in the term to be explained as part of the explanation, and we therefore reject the explanation as invalid. But we do not therefore reject the notion as erroneous, but accept it as something ultimate, which, while it does not admit of explanation, does not require it.
>
> But this does not apply here. An idea may be valid of reality though it does not admit of valid explanations. But it cannot be valid of reality if its application to reality involves a contradiction. Now we began by pointing out that there was such a contradiction in the case of time—that the characteristics of the A series are mutually incompatible and yet all true of every term. Unless this contradiction is removed, the idea of time must be rejected as invalid. It was to remove this contradiction that the explanation was suggested that the characteristics belong to the terms successively. When this explanation failed as being circular, the contradiction remained unremoved, and the idea of time must be rejected, not because it cannot be explained, but because the contradiction cannot be removed.

Now I think McTaggart is wrong in what he says in this last paragraph for exactly the same reasons for which Parmenides was wrong in the previous case. The charge of circularity can, and should, be turned back against him. The "incompatible determinations", being future, being present, and being past, are more immediately related to time than nonexistence and existence in the Parmenidean case, this may perhaps lend some additional appeal to McTaggart's

claim that whoever thinks that these determinations are not incompatible, because they are possessed by events not simultaneously but successively, is thinking circularly. But this is just a rethorical advantage. McTaggart's case is not really stronger than Parmenides's.

For it is not the case that McTaggart's objector is trying to remove a contradiction in the idea of the A series employing a circular explanation. Just like in the Parmenidean case, the situation is much more fairly characterised as one in which the objector points out that there is no contradiction to start with in the idea, unless we assume that what the idea is about is unreal. There is no contradiction in the concept of the A series unless we assume that time is unreal. It has not been proved, to use McTaggart's own words, that the idea of the A series involves a contradiction and so it cannot be valid of reality. All that has been shown is that *if it is not valid of reality, then it involves a contradiction*. This is an interesting conclusion, but it falls short of what McTaggart wanted to prove. This is a conditional that can be true if both its antecedent and its consequent are false. And McTaggart wanted to prove the consequent. If McTaggart had an independent argument in support of the truth of the antecedent, then of course he could prove the consequent. But McTaggart can prove that the antecedent is true only on the assumption that the consequent is true. So his argument is circular.

McTaggart offers also another way of presenting what he thinks is wrong with what his objector is saying. In this version, instead of a vicious circle, he mentions a vicious regress.

> If we avoid the incompatibility of the three characteristics by asserting that M is present, has been future, and will be past, we are constructing a second A series, within which the first falls, in the same way in which events fall within the first. It may be doubted whether any intelligible meaning can be given to the assertion that time is in time. But, in any case, the second A series will suffer from the same difficulty as the first, which can only be removed by placing it inside a third A series. The same principle will place the third inside a fourth, and so on without end. You can never get rid of the contradiction, for, by the act of removing it from what is to be explained, you produce it

> over again in the explanation. And so the explanation is invalid.

I think it doesn't improve McTaggart's case at all. First of all, I don't see that we are making "the assertion that time is in time." We are making, rather, the assertion that the temporal determinations of the A series, future, present and past, apply at certain times and does not apply at others. But there is nothing surprising in it if we consider how we get them from the temporal determinations of the B series, earlier and later, which McTaggart considers real. If we supplement earlier and later with a third determination, simultaneous, we will find it that the one-place predicates of the A series can be obtained from the two-place predicates of the B series by holding one of the arguments "fixed". Not fixed in the sense that it would always be the same event. But it preserves an identity, though; it is the *now* of consciousness. What is past is earlier than it, what is present is simultaneous with it, and what is future is later than it. Since the *now* of consciousness is moving with time, the determinations we got from the determinations of the B series by substituting a moving point of reference in the place of one of their arguments, will apply to a certain event at certain times, and not at others. There is nothing mind-blowing in it, as McTaggart seems to gesture.

As far as the "regress" is concerned, I think it is as useless, from McTaggart's point of view, as the "circle" was. It is not the case, as McTaggart claims, that his objector "can never get rid of the contradiction, for, by the act of removing it from what is to be explained, [he] produce[s] it over again in the explanation". There has never been a contradiction, unless we assume, what needs to be proved by McTaggart, that time is unreal. And the objector is not attempting to "explain" how is the idea of the A series coherent. If he is "explaining" anything, it is how McTaggart is wrong. McTaggart is wrong presupposing the conclusion he wants to reach, and thereby creating a "contradiction" in the idea of the A series, by an appeal to which he will argue for his conclusion. The objector is just pointing out this circularity in McTaggart's thought, and the ascent to ever higher order determinations, future in the past, present in the present, past in the future, etc., is completely useless, from McTaggart's perspective, since he won't get a problem with these higher order determinations either, unless he presupposes the conclusion, which he wants to establish on the basis of it.

So I think we can safely conclude that this twentieth century Eleatic argument falls way short of convincing us of the unreality of becoming.

The argument from the special theory of relativity against becoming and presentism

But there is an apparently much stronger argument, based on a very successful modern scientific theory, the special theory of relativity (STR), which, *prima facie*, seems to prove that becoming is unreal.

The argument from STR to the unreality of becoming has two parts. One is an argument to the effect that an ontological doctrine about time called *presentism* is essential to the idea of becoming, and the other is an argument aiming to show that presentism is untenable if STR is accepted, since STR proves that the idea of the *present*, on which presentism rests, does not apply to anything real.

It is plain that the idea of the present is indeed constitutive of the idea of the A series, and that of becoming. If there is something wrong, conceptually, with the notion of the present, then the problem is inherited by the idea of the A series, and the concept of becoming. It is a minimal condition for the reality of the temporal determinations of the A series, and for the reality of becoming, that the present be real.

What is so special about the present?

The answer is that the present is special ontologically. An A-theorist of time believes that the temporal determinations of the A series correspond to ontological determinations. What is future is not real yet. What is past is not real any longer. Yet, there is an asymmetry between these two modes of nonexistence. About a future event that is not real yet there may be an ambiguity. As far as its ontological status is concerned, it may turn out more than one way. This is why some future events may be up to us. There is no similar ambiguity about past events. They are what they are, because they have been real, and when they were real, their identity consolidated. This is becoming. They became what they are unchangeably when they were real. But when were they real? If not when they were future, when they were not yet real, and not when they were already past, when they were not real any longer, then the only remaining alternative is that they were real when they were present. Becoming takes place in

the present. The present is special, because it is the locus of real existence.

Any philosopher who is an A-theorist of time, and who believes in becoming, must believe in something along these lines about the ontological exclusivity of the present. This is presentism: the doctrine that, in some perfectly good sense, only what is present is real, or, what is the same, that the present is real in a sense in which the future and the past aren't.

Many philosophers have found presentism puzzling. And many of them found it puzzling because of the peculiar spatio-temporal extendedness of the present, namely that it has no extension along the time axis, while it is extended spatially. The worry related to the temporal unextendedness of the supposed locus of real existence is very old, it was famously articulated by Augustine. The worry related to its spatial extendedness is relatively new, it is a consequence of the rise of the special theory of relativity, and the first philosopher who gave it a precise formulation was probably Kurt Gödel.

Although Augustine is sometimes counted with the presentists, my impression of him is that he would have only been a presentist if he had believed in the reality of time, but he hadn't. In the *Confessions* he treats presentism as the only way to be realist about time, but he seems to think he can show that presentism leads to absurdities. The point of the whole discussion in Book XI seems to be that time is, after all, just a mental thing, lacking objective reality. So, in my impression, Augustine is pretty much in the same boat with McTaggart.

The source of Augustine's puzzlement about presentism is that he thought the present must be lacking any extension:

> If any portion of time is conceived, which cannot now be divided into even the minutest particles of moments, that alone is what may be called present. And that flies by with such speed from future to past that it cannot be lengthened out in the least, for if it is extended, it is divided between past and future. The present has no extension or length.[189]

[189] Augustine 1986, p. 243

If we combine this thesis with the thesis of the nonreality of the past and the future, then, in Augustine's view, there arises a problem with the measurement of time. Augustine says he knows we measure temporal distances, periods, intervals in time, but it is unclear to him how. For, if the past and the future are nonexistent, then, he says, it is not very likely that we can measure periods in the past or in the future.

> But past times, which no longer are, or future times, which are not yet, who can measure? Unless, perhaps, anyone would dare to say that what is not can be measured.[190]

But then there aren't many options left. Augustine considers the possibility that we measure time "when it is passing", i.e. in the present, but he finds himself confronted with the problem that, if the present is extensionless, then it surely cannot contain the intervals that we measure.

> But how do we measure present time, since it has no extension? It is measured while it passes; but when it shall have passed, it is not measured; for there will not be aught that can be measured. But whence, in what way, and whither does it pass while it is being measured? Whence, but from the future? Which way, save through the present? Whither, but into the past? From that, therefore, which as yet is not, through that which has no extension, into that which now is not.[191]

Now, as far as the measurement of time is concerned, I do not share Augustine's worry.

If presentism is true, then change is real. If change is real, then we can compare periods of time to processes of change, the rate of which we may agree to consider as standard. We may call these standard processes of change clocks. This is how time is measured.

---

[190] p. 244.

[191] With this, said in Chapter 21 of Book XI, Augustine did not make a dramatically new point, however. Aristotle said practically the same in Book 4, Chapter 10 of *Physics*, about eight centuries earlier. Just like extensionless points do not add up to form an extended line of whatever minute length, extensionless moments, no matter how many of them, will never constitute a period.

Yes, we are measuring intervals that are (mostly, that is, apart from their endpoint) in the past. But there is nothing wrong with that, because we are doing it by comparing them to standard processes of change that are also (mostly, apart from their endpoint) in the past. The comparison is based on simultaneity. It is based on the simultaneity of the beginning of the measured interval with a certain stage of the standard process of change to which it is compared (e.g. the hands of a clock being in a certain position), and the simultaneity of the last moment of the measured interval with another certain stage of the same standard process. Simultaneity is always checked when it is present. So, in this sense, we perform the measurement in the present. But, again, there is nothing wrong with that, because the simultaneity of two events is a state of affairs that can obtain in the extensionless present.

So I think we may conclude that there is nothing incomprehensible in the measurement of time, even if it is accepted that only the present is real, and that the present is extensionless.

But I think there is a deeper problem with the combination of presentism with the thesis of extensionless present. The problem arises because the vanishingly thin present ought to contain all that is real, and it probably cannot.

Consider, for example, a musical note as it sounds. Sound is a mechanical wave. The pitch of the note is determined by the frequency of the wave. Does the frequency of the wave exist in the extensionless present? It seems that it doesn't. It would be a contradiction in terms to suppose so. But if it does not exist in extensionless present ever, then it seems to follow from presentism that the pitch of musical notes is something that can never really exist. But we seem to be phenomenally aware of sounds of different pitches. We are aware of them with our phenomenal perception directed at the present.

One possible explanation to this phenomenon could be that phenomenal sounds are the creations of the mind, and our minds rely not only on immediate perception, but on memory, too, when they are generating acoustic sensations.[192] But this explanation is really hard to believe. There seems to be an immense difference between listening to music and remembering music. In our memory we store the sequence of pitches, but that the very pitch that is presently being

[192] Augustine is giving this explanation to the phenomena of melody and rythm in Chapters 26 and 27.

heard would be remembered rather than perceived strikes me as a fantastic idea.

But maybe the idea of the *specious present* comes to our rescue at this point. The idea came from a psychologist, William James, more than a hundred years ago.[193] It holds that the mind has a tendency to experience a number of events happening at one time, although, in reality, they are spanning an interval of finite extension. Of this short interval we are "immediately and incessantly sensible", and it serves as "a prototype of all conceived times", as James put it.

Now if the doctrine of the specious present is true, which I believe is a matter of empirical psychology, it asserts that *the now of consciousness* is extended, at least in some of its aspect. But it seems that there are reasons that suggest that the *objective present*, i.e. the locus of real existence, must be extended, as well.

To give a familiar example, I think the problem of an extensionless present as the exclusive locus of real existence is at the heart of also Zeno's famous paradox of the arrow.[194]

What Zeno seems to have in mind is this: Throughout its flight, at every instant the arrow fills out a definite space, corresponding to its length. As far as an extensionless instant is concerned, filling out a definite space is indistinguishable from being at rest at that place. So the arrow is at rest at every particular instant. So the arrow is at rest at every instant throughout its flight, which is a contradiction. So we deduced a contradiction from the supposition that the arrow moves. So, contrary to appearance, we have to discard the hypothesis that the arrow ever moves.

I think this argument of Zeno's draws on, not just a hidden premise, but on a whole hidden argument. This argument would go like this. The only alternative to the Eleatic view of timeless existence is presentism. A presentist must hold that if there is a matter of fact to distinguish between two situations that might obtain at an instant, then this fact must hold within that instant, since nothing really exists apart from what is contained in that instant. There is no fact contained in an instant that would distinguish between an arrow being at one place being in motion, and the same arrow at the same place being at rest. That is why they are indistinguishable.

With this supplementary argument in view, the situation seems to be that one has to give up either motion or presentism. And if

---

[193] James 1890

[194] As reconstructed from Aristotle's *Physics* 239b5-7.

presentism is really the only alternative to the Eleatic view of the world, then the latter prevails either way.

If, however, there is a way to account for the reality of things that are not contained in the present moment without going Eleatic about existence, then Zeno's argument can be sinked, because then an arrow being at one place at rest, and the same arrow being at the same place in motion are distinguishable.

To be more general, the way physical science works seems to suggest that a full description of the physical state of a system involves not only the values of the physical properties that characterize the system at a given instant, but also their rate of change. A frozen picture, that is, the information an extensionless present may contain of a physical system, underdetermines its physical state. The information contained by an extensionless instant and the laws of physics together fail to determine how the system is to evolve, even under classical assumptions. The most obvious example is the momentum. A frozen picture of the world contains no information of the momentum of physical objects. Yet, to know how a physical object will move on, one has to know its momentum. Generally, the initial conditions, that are required for the equations that express the laws of physics to produce a unique solution, need to involve the first derivatives of co-ordinates that describe the system, at a given time, as well as the co-ordinates themselves. Now the first derivative of a physical property at a given time is not a piece of reality that can be contained in an extensionless instant. It is the rate of the change of the property with time. Mathematically it is construed as the limit of a series of ratios of differences of the values taken by the property in question at different times, and of the differences of those times. It could not be real, it seems to me, if only an extensionless present was real. The extension that needs to be real for a limit of such a series to be real can be infinitesimally small, but it has to be finite, i.e. non-zero.

If it is true that an extentionless present cannot accommodate everything that is physically real at that particular time, then maybe objective present is specious just like psychological present is. Or, there is an even simpler solution to this problem a friend of becoming may choose. The unreality of the past is not essential for becoming to be real. What is essential is the unreality of the future. So a theorist of becoming may relax the ontological exclusivity attributed to the present by presentism, and allow the past to be real, as well. (Not

necessarily in exactly the same sense in which the present is real, but certainly in a sense that sharply distinguishes the past ontologically from the future, and not so sharply from the present, so that the theorist of becoming be able to account for the physical reality of the momentum, and similar physical properties. This would allow for distinguishing between the resting and the flying arrow at every instant.) The present remains the locus of coming to be, and the future remains as it was under presentism, as one might put it, not actual, only possible. Hence the name for this more relaxed version of presentism: possibilism. If we are possibilists, rather than presentists, then we need not be worried if it is the case that an extensionless instant cannot contain everything that is real at the given time.

Yet the present is at the heart of this more relaxed ontological doctrine just as much as it was at the heart of presentism. To maintain an ontological distinction between the past and the future, which seems to be a minimal condition for becoming to be real, the dividing line, or, rather, the dividing three-dimensional hypersurface of spacetime, between the past and the future must be objective.

Now the special theory of relativity entails that it isn't.[195]

On either presentism or possibilism (I lump the two together under the label of presentism from now on), we naturally think that what divides between what has not yet, and what has already, become real, is what the three-dimensional space is filled out *now*. It is very much like what we see when we look around (say, from a high place). It is not quite that, though, since, strictly speaking, we never see the present, what we see is always the past (more precisely, we see a part of the surface of our past lightcone[196]). The further we look into

---

[195] Other possible objections to presentism, which are not concerned with either the temporal, or the spatial extendedness of the present, are discussed and answered in Ned Markosian's contribution to the relatively recent *Oxford Studies in Metaphysics* (Zimmerman 2004). These include the objection that if presentism was true, there would be no singular propositions, having a truth-value, about past and future objects, because their having a truth-value depends on the existence of the object they are about, and a closely related one, endorsed by Quine (1987), asserting that, if presentism was true, we could not stand in any relation to any non-present object, for the reason that one of the relata would simply not exist. These worries, to my mind, are satisfactorily answered by Markosian. (I am dissatisfied, however, by his treatment of the objection taken from the special theory of relativity.)

[196] Two events are lightlike separated if they could be two points of a lightray, that is, if the ratio of their spatial and temporal separation is the speed of light. If we represent the spacetime points that are lightlike separated from a point of reference on a diagram, with

space, the further we look into the past, due to the finiteness of the speed of light. But, unless we look up at the sky on a clear night, we see pretty much the same as if we could look into the present, since the speed of light is really big compared to the rates of the changes we normally observe around ourselves. Steven Savitt speculates that this is the explanation for the formation of the idea of *a spatially extended now* in the course of the evolution of the basic conceptual furniture of human cognition.[197]

Now the present so conceived is supposed to be a slice of space-time, a three-dimensional snapshot of the universe, going from one end of it to the other, containing whatever there is, and happens, simultaneously with the here-and-now of our consciousness. According to STR, however, only the *now* of the *here-and-now* is unproblematic; simultaneity between distant spacetime points is relative to the choice of a frame of reference. The *there-and-then* of distant events decomposes to *there* and *then* differently for inertial observers that are in motion relative to each other.

Or, to put it more formally, whereas in Newtonian spacetime the temporal distance of any two events is a well-defined, objective measure (observer-independent, or, what is the same, independent of the choice of the frame of reference), this is not so in the special theory of relativity. In the geometry of STR's Minkowski spacetime, only a measure which is characteristic of both the spatial and temporal separation of two events, taken together, is objective (in the sense that it is observer-independent), the spatial and the temporal separation of two events become well-defined, one by one, only when a frame of reference is introduced, or, what is the same, spatial and temporal separation disentangled from each other exist only from the perspective of an observer.

This result leads into insurmountable difficulties when we are trying to identify points of Minkowski spacetime that are

---

one spatial dimension suppressed, they are on the surface of two cones facing each-other with their apexes at the point of reference. The apex of the lightcones is the point representing the present moment of the worldline of the observer whose co-ordinates are represented on the diagram. Events inside either of the two lightcones are said to be timelike separated from the point of reference. Events outside both lightcones are spacelike separated from the point of reference. On the special theory of relativity, involving the thesis of the invariance of the speed of light, the lightcone structure is invariant, i.e. it is something on which different inertial observers in relative motion whose worldlines intersect at the point of reference can agree.

[197] Savitt forthcoming, pp. 15-6.

simultaneous, i.e. when we are trying to identify events that are separate spatially but not separate temporally. The present is supposed to be one such simultaneity plane of Minkowski spacetime. This is what undermines the objectivity of the present.

There is a spacelike three-dimensional hypersurface of spacetime that is the present for me. Suppose there is another observer who moves relative to me. There is a point of his worldline which is part of my present. When he is in that point, he is present to me. But if we take the spacelike three-dimensional hyperplane of spacetime that is the present for him when he is in that point of his worldline, it turns out to be different from my present. Someone's present, who is present to me, is different from my present, if we are in motion relative to each other. Some events that are yet to take place in my frame of reference are already past in his.[198]

Now what is the truth about these events? What is their ontological status? Have they already become real? In his frame of reference it looks as if they had. But in mine it looks as if they had not. Why would my truth about them be any better, or any worse, than his? We are just two physically equivalent observers in relative motion. But if the present is to divide between two ontological categories, then it should be objective. Neither his present nor mine (or anybody's) is a good candidate for the title of the objective present, for they are not the same, and it is hard to see what would be so special about me, or him, or anybody who's present would

---

[198] It is interesting to see how little it affects the intersubjectivity of our perceived, psychological present, under normal circumstances. Savitt, when accounting for the formation of the idea of an intersubjective and spatially extended present, presents an interesting calculation about the differences of the two presents of two people walking past each other (ibid). He says he was informed by experimental psychologists that the duration of specious present varies (inter- and intrapersonally) between .5 and 3 seconds. For the sake of simplicity Savitt takes it to be 1 second. The volume of a specious present he takes to be the part of spacetime with which we can exchange causal signals in this 1 second. He proposes that we should consider the intersection of the part of spacetime on which we may potentially have a causal influence after the first moment of this 1 second long period, and the part of spacetime that can causally influence us until the last moment of the same period. (This is the intersection of the future lightcone of the first moment and the past lightcone of the last moment, along the worldlines of the two observers.) Now, Savitt continues, "suppose that you walk past me at a reasonable pace of 4 km/hour, that we call our meeting e, and that we compare the volumes of your present and my present, assuming they are symmetric about e (that is, each present extends .5 seconds to the future and to the past along our two world lines). Then our two presents agree—that is, include the same events—up to about one half of one millionth of one percent ($\sim 5 \times 10^{-9}$)."

supposed to be the real, ontologically significant present, if we are physically equivalent, and no observer has any privilege over the others.[199]

Kurt Gödel summed up the situation like this:

> The existence of an objective lapse of time…means (or, at least, is equivalent to the fact) that reality consists of an infinity of layers of "now" which come into existence successively. But, if simultaneity is something relative in the sense just explained, reality cannot be split up into such layers in an objectively determined way. Each observer has his own set of "nows", and none of these various systems of layers can claim the prerogative of representing the objective lapse of time.[200]

Now someone might come up with the idea that, if no one's present is better than anybody else's, why don't we use a democratized notion of the present.[201] The notion of the present rests

---

[199] Roger Penrose (1989) illustrates the problem this result poses for an advocate of an objective and ontologically significant extended present by asking us to consider the difference between the presents of two earthly observers at such a distant place as the Andromeda. Penrose's two observers, Alice and Bob walk past each other at exactly the same relative speed as in Savitt's above example, 4 km/hour (see the previous footnote). Despite the insignificance of the difference between their psychological presents, i.e. the parts of spacetime with which they can exchange causal signals within the length of their specious presents, as defined by Savitt, at the distance of the Andromeda (about two million lightyears) their simultaneity-planes come quite significantly apart. If, for the sake of simplicity, we assume that the Andromeda is at rest with respect to Earth, and that Alice walks towards the Andromeda, whereas Bob walks away from it, then their simultaneity planes intersect with Andromeda's worldline 5 ¾ days apart. Penrose asks us to suppose that something really significant happens in these 5 ¾ days. The Andromedeans launch a fleet to invade Earth. (The argument seems to be designed to be the relativistic analogue of the Sea Battle Argument for fatalism in Aristotle's *De interpretatione*). This event is in Alice's past and in Bob's future, even though they walk past each other right now on the street. Now it is hard to imagine that there can be an ambiguity about the ontological fixity of this event, says Penrose: "Two people pass each other on the street; and according to one of the two people, an Andromedean space fleet has already set off on its journey, while to the other, the decision as to whether or not the journey will actually take place has not yet been made. How can there still be some uncertainty as to the outcome of that decision?" Penrose takes it as an argument for the nonexistence of any uncertainty about any decision, or any event, ever. "If to either person the decision has already been made, then surely there cannot be any uncertainty. The launching of the space fleet is an inevitability." (p. 303.)

[200] Gödel 1949a, p. 557. Cited by Savitt, forthcoming.

[201] Putnam has considered this idea (1967).

on the notion of simultaneity, and it was the observer-relative nature of simultaneity that caused the problem. What if we tried to democratize simultaneity, by replacing observer-relative simultaneity with simultaneity-in-one-or-another-inertial-observer's-frame-of-reference? Present so democratized would be the collection of events that are simultaneous, with the event that is here and now, in one or another observer's frame of reference.

The bad news is that the smallest non-empty subset of Minkowski spacetime that is closed under this democratized relation "simultaneous in one or another inertial observer's frame of reference" is the Minkowski spacetime itself in its entirety.[202] So, on a democratized account of the present, which would be based on the democratized version of simultaneity, all events of the entire "history" of the universe would equally be present, and the crucial ontological difference between what is already existent and thereby definite, and what is not, would fall.

So it seems that, on the special theory of relativity, presentism cannot be maintained, and thus the reality of becoming cannot be maintained either.

The first generation of relativistically minded physicists already appreciated this problem. In an often quoted passage of his autobiography, Carnap recalls a conversation he had with Einstein in the early 1950's, which concerns the conflict between relativity and presentism.

> Once Einstein said that the problem of the Now worried him seriously. He explained that the experience of the Now means something special for man, something essentially different from the past and the future, but that this important difference does not and cannot occur within physics. That the experience cannot be grasped by science seemed to him a painful but inevitable resignation.[203]

Herman Weyl once made a similar comment that concerned becoming directly:

> The objective world simply is, it does not happen. Only to the gaze of my consciousness, crawling upward along the

[202] Putnam, ibid.
[203] Carnap 1963, p. 37.

> life line of my body, does a section of this world come to life as a fleeting image in space which continuously changes in time.[204]

To my knowledge, the first philosopher to discuss the bearing of STR on the metaphysics of time was the great Kant scholar Ernst Cassirer.[205] Both he and Gödel thought that the observer-relative nature of the present in relativity theory entailed that time was *transcendental* in Kant's sense.[206] Indeed, Weyl's above comment comes very close to what Kant wrote in the *Prolegomena* and in the *Critique of Pure Reason*, that time (and space) are

> determinations adhering not to things in themselves, but to their relation to our sensibility[207],

and that

> [t]hose affections which we represent to ourselves as changes, in beings with other forms of cognition would give rise to a perception in which the idea of time, and therefore also of change would not occur at all.[208]

It should be mentioned that if we go Kantian about time, we have also the possibility of going Kantian about libertarian freedom as well. Kant was a libertarian and a nonrealist about becoming. The unreality of becoming is not strictly speaking incompatible with libertarian freedom. It is not inconceivable that we exercise libertarian freedom timelessly, as it was discussed in the second chapter. (This was the

---

[204] Weyl 1949, p. 116.

[205] Cassirer 1920, cf. Dorato 2002.

[206] Meaning that time is not part of the mind-independent furniture of reality, rather, it is an a priori intuition of the mind, with the help of which it organizes sensory data. Gödel wrote that "the agreement described between certain consequences of modern physics and a doctrine that Kant set up 150 years ago in contradiction both to common sense and to the physicists and philosophers of his time, is greatly surprising, and it is hard to understand why so little attention is being paid to it in philosophical discussion of relativity theory" (Gödel 1990, vol. 2, p. 236). A present day advocate of the view that the metaphysical status of Minkowski spacetime is best interpreted the Kantian way is Mauro Dorato (2002).

[207] Kant 1783/1953, § 11, p. 36.

[208] Kant 1787/1970, I. Transcendental Doctrine of Elements, First Part. Transcendental Aesthetic, Section 2. Time, p. 79, cited by Gödel 1949a, p. 558, cited by Dorato 2002.

only way Kant saw to reconcile freedom and moral responsibility with determinism, which he also held true—of events that follow each other by natural necessity in the phenomenal flow of time—given that he found the Humean compatibilist conception of freedom utterly unsatisfactory.) This move, however, endangers rational decision-making on the ground of past experience, as it was also discussed in the second chapter, so I think libertarian theorists should seek other ways of saving libertarian freedom, as long as it is possible.

Before we join with the above mentioned great authorities, and settle on discarding realism about the flow of time, and so about becoming, it is important to see whether we are really bound by empirical data to do so.

Ways of trying to resist the relativity of simultaneity

The feature of the special theory of relativity from which the nonexistence of an objective present is derived is the relativity of simultaneity. The relativity of simultaneity follows from the way the simultaneity of spatially separate events is defined in STR.

The simultaneity of two distant events is determined by local clocks. If there are clocks near both of them, and they read the same time when the two events happen, then the two events are simultaneous. Provided, of course, that the two clocks have previously been synchronized.

The invariance and isotropy of the speed of light[209], i.e. that the speed at which light travels is the same in all inertial frames of reference in all directions[210], provides us with a method of synchronizing distant clocks. Consider two clocks, A and B, at rest relative to each other, at some distance from each other. Clock A and clock B are synchronous if, and only if, it is true that if a light-signal is emitted from the locus of clock A when the time read on clock A is t_0, and the time read on clock B when the light-signal is reflected back

[209] It is often claimed to have been the finding of the famous Michelson-Morley experiment, which was originally designed to determine the speed at which Earth travels in the Aether, i.e. relative to the absolutely resting frame of reference, on the basis of the shift in the observable speed of light which the movement of Earth relative to the Aether was supposed to cause. The result of the experiment was that there was no such shift. However, on a closer look, the experiment testifies only that the two-way (round trip average) speed of light is invariant, it says nothing about the one-way speed of light. More would be said about it a little later.

[210] Inertial frames of reference may move relative to each other at a constant speed.

from its locus is t_1, then the time read on clock A when the light signal arrives back to its locus is $2t_1-t_0$.[211] This definition means that along the worldline of clock A the event that is halfway between the emission and the return of the-light signal is taken to be simultaneous with the event of its reflection from the locus of clock B. If the distance of clock B from clock A is x, and the speed of light is c, then $t_1 = t_0+x/c$.

Now from this definition, which is in effect in all inertial frames of reference, the relativity of simultaneity immediately follows. Consider the thought experiment Einstein himself devised to illustrate it. There is a train passing by a platform at a constant speed. The platform is exactly as long as the train. There is an observer in the middle of the train, and another in the middle of the platform. They correspond to two frames of reference that are in uniform motion relative to each other. Suppose that a light flashes when and where the two observers pass each other. Some of this light will travel toward the front of the train, some toward the back of the train. Since distant clocks were synchronized with light-signals in the frame of reference attached to the observer on the train, the clocks of this frame of reference will read the same time when the two lightrays emitted from the middle reach the front and the back of the train. So these two events will be simultaneous in this frame of reference. In the frame of reference attached to the observer standing in the middle of the platform, however, the events of the two rays of light hitting the two ends of the platform will be simultaneous. The event when the lightray travelling backwards relative to the motion of the train hits the back of the train will be earlier than this event, because meanwhile the end of the train will have covered some distance from the back end of the platform toward the place from where the lightray was emitted. The event when the lightray travelling forwards relative to the motion of the train hits the front of the train is later than this event, because meanwhile the front of the train will cover some distance moving away from the front end of the platform. So the event of the backward moving lightray's hitting the back of the train, and the event of the forward moving lightray's hitting the front of the train will be simultaneous in one frame of reference, and will not be simultaneous in the other. This shows that the two frames have different simultaneity planes. Their simultaneity planes that contain

[211] Einstein 1905.

the moment when the two observers pass each other, and the light is emitted, will be different too. To be vivid, the *now* of two observes who pass each other *here and now*, consisting of the events that happen when the local clocks of their frame of reference read the same time as the clock *here and now*, are different, if they are in relative motion.

Now there are several ways that suggest themselves to resist this conclusion.

Siding with Lorentz and Fitzgerald

One is to suppose that although the events of the lightrays hitting the two ends of the train in the example are simultaneous according to one observer, and nonsimultaneous according to the other, it might be the case that one of them is right, and the other is wrong. (Or that both of them are wrong, and a third observer is right. The important thing is that there is an observer who is right about the question of their simultaneity, and those who disagree are wrong, i.e. that there is a truth about this matter.)

Some might suspect, instinctively, that the observer on the platform is right. But of course, if there is an Aether, and the train happens to run at the same speed relative to the platform as the part of the surface of Earth to which the platform belongs travels relative to the Aether, but in the opposite direction, then, we might probably say, the observer on the train is right.

There is a strong physical symmetry of inertial frames in relative motion: the laws of physics are expressed by exactly the same mathematical equations in the co-ordinates of both frames of reference, and there is no physical experiment that would determine which of the two frames is truly in motion, relative to the frame of reference, if there is such a frame, that is at absolute rest, and also the speed of light will be measured to be the same in all of them, in all directions.

On the other hand, the spatiotemporal distances between events disentangle into spatial and temporal distances only after a frame of reference has been chosen. The measuring rods and clocks of different frames of reference travelling at a non-zero speed relative to each other will measure different spatial and temporal separations for the same couples of events.

The special theory of relativity declares that among the physically equivalent inertial frames of reference there is none which would

have more claim on capturing true spatial distances and true periods of time that separate the two events, than any other.

This principle, I believe, is best understood as a metaphysical hypothesis. It is not only, the principle says, that there is *no knowing* of which frame captures true space and true time, because the physical equivalence of inertial frames blocks our epistemic access to that fact, it is that there is *no such fact*, metaphysically speaking. Frankly, the principle says that we should not suspect a metaphysical asymmetry where there is no physical asymmetry.

But, of course, empirical physics does not ground such a principle. Empirical physics tells us only that there is no telling which frame is at absolute rest. It does not follow from it that there is no such property as being at absolute rest. Empirical physics, in combination with Occam's simplicity principle, or with a verificationist inclination to conflate the epistemological issue with the metaphysical one, may motivate us to discard the idea of an absolute space and time, but it does not dictate it, strictly speaking. If our metaphysical view of the world, for example, requires us to hypothesize an absolute space and an absolute time, physics does not rule it out. If it was the case that phenomena such as becoming, and the asymmetry between the past and the future, can be accounted for only on the hypothesis of an absolute space and time, the empirical facts that ground the special theory of relativity would not prohibit us from being realists about these phenomena.[212]

George Fitzgerald[213] and Hendrik Lorentz[214] proposed an account of the 'null-result' of the Michelson-Morley experiment which is consistent with the hypothesis of an absolute space and time, and contradicts the principles of special relativity, i.e. the metaphysical equivalence of inertial observers and the invariance of the (real) speed of light.

Take a frame of reference, say Lorentz and Fitzgerald, in which the Maxwellian equations of electrodynamics are valid, and consider a charged point-like particle. If it is at rest relative to this frame of reference, then its electric field is spherically symmetrical. If, however, it is in uniform motion, then its field undergoes a deformation. It can be calculated from the Maxwell equations that it undergoes a contraction in the direction of the movement. Now suppose that the

---

[212] For a thorough discussion of the verificationism involved in STR see Smith 1998.
[213] 1889.
[214] 1892.

moving pointlike particle is the nucleus of an atom. If there is an electron circling around it, then the orbit of the electron around the nucleus is suppressed, following the contraction of the nucleus's electric field, by a factor of $\sqrt{(1-v^2/c^2)}$. Now suppose that it is a useful model of objects made of atomic matter, such as the arms of a Michelson-Morley interferometer, or a measuring rod, or a clock, for example, that they consist of nuclei, and electrons moving in the electrostatic fields of the nuclei (and each other), and their lengths are determined by the space occupied by their electrons. If the orbits of their electrons contract according to the Maxwell equations, then these objects are expected to undergo a corresponding contraction in the direction of their movement. Lorentz and Fitzgerald pointed out that if the arm of the Michelson-Morley interferometer that points in the direction of Earth's movement in the Aether contracts to a length of $\sqrt{(1-v^2/c^2)}$ times its original length, as predicted by this simple model of atomic objects, then it accounts for the null-result of the experiment, if it is assumed, as we would assume it classically, that the speed of light is c only relative to the Aether, and it is $c \pm v$ relative to the arm of the interferometer ($c-v$ on the way from one end to the other, $c+v$ on the way back).

It can also be calculated that an observer travelling with the interferometer will observe nothing of this change in the speed of light. His measuring rods will contract to a length of $\sqrt{(1-v^2/c^2)}$ times their original lengths in the direction of the movement of the whole laboratory, and, since his clocks deform, too, they slow down, and, in consequence, between any two events resting clocks will read $1/\sqrt{(1-v^2/c^2)}$ times more time than the moving ones. Lorentz has pointed out that if we choose to express spatial distances and periods of time in the co-ordinates of the moving frame of reference, i.e. in terms of distances measured by contracted rods, and the lengths of temporal intervals as read on deformed clocks, then in these new co-ordinates the laws of physics that account for physical phenomena in an empirically correct way will look exactly as if the whole system was at rest, that is, they will have the same mathematical form as the laws of physics that account for physical phenomena in an empirically correct way in the resting frame of reference, expressed in the co-ordinates measured by undeformed rods and clocks. (E.g., it will seem to the physicist in the moving laboratory that the Maxwellian laws of electromagnetism hold in his laboratory, even though they actually don't. They seem to hold because every measurement in the moving

laboratory is performed with deformed instruments. The electrostatic field of a point-particle that is at rest relative to the laboratory, for example, seems spherically symmetrical, although, in reality, i.e. in the co-ordinates of the resting frame (the true co-ordinates measured by undeformed devices) it is contracted in the direction of the movement of the laboratory.) Therefore, there can be no experiment a physicist in a moving laboratory could perform, the result of which could inform him about the metaphysical fact that he and his laboratory are moving. (E.g., expressed in these new co-ordinates, that the moving observer measures by his deformed devices, the speed of light will be c, just as it is in the resting frame of reference, expressed in terms of co-ordinates measured by undeformed devices, so the moving physicist have no chance to perform a measurement that would inform him about the fact that the speed of light relative to his laboratory is different.)

So Fitzgerald and Lorentz explained why the speed of light *appears to be* invariant, without giving up absolute space and absolute time, and without giving up the idea that light travels at speed c, calculable from the Maxwell equations, only if distances and periods are expressed in the co-ordinates of absolute space and absolute time. It is sometimes said that they did so by introducing an *ad hoc* hypothesis, i.e. the contraction of moving objects in the direction of their movement, but this is not correct. The Lorentz-Fitzgerald contraction of moving objects is not an *ad hoc* hypothesis, rather, it is entailed by the Maxwellian electrodynamics of moving bodies, if it is accepted that atomic matter is held together by electrostatic forces.

From the two theories, Lorentz's and Fitzgerald's on the one hand, and Einstein's, on the other, it is impossible to choose empirically. All their testable predictions are the same. Lorentz contraction and time dilation are the same, and equally real, on both theories. According to the special theory of relativity, measuring rods and clocks are four-dimensional objects, different observers see different collections of points of these four-dimensional objects as simultaneous, or as staying at rest for a period of time, and this accounts for the differences in their lengths, and in the lengths of the periods measured by them. Sometimes it is said that on STR Lorentz contraction and time dilation do not require a specific explanation based on a physical model of rods and clocks. This is true. However, the deformation of electric fields, on which the Fitzgerald-Lorentz explanation is based, is there, and it is equally significant.

Apart from their view of space and time, there is no difference between the two theories whatsoever. The only difference between them is that Einstein proposed to call "space" and "time" also the co-ordinates of moving frames, which, he admits, are measured by deformed measuring devices, whereas Fitzgerald and Lorentz insist that only the spatial and temporal co-ordinates measured by undeformed devices are properly called "space" and "time", although they admit, that there is no knowing of which devices are undeformed. From a physical point of view, this is only a terminological disagreement. From a metaphysical point of view, however, the difference between these two physically equivalent theories is enormous, because of the ontological significance with which our notion of time is burdened.

So one way to resist the Gödelian-Putnamian conclusion about the unreality of becoming is to side with Fitzgerald and Lorentz in the debate between them and the special theory of relativity. One can do that without running the risk of being disproved by physics. Although it is not very fashionable, the Lorentz-Fitzgerald view has always had, and still has, prominent supporters among both physicists and philosophers.[215] Some of them hold that, in the context of quantum mechanics, the non-locality arising from the correlatedness of the quantum states of spacelike separated objects with a common past, like those described in the family of Einstein-Podolsky-Rosen-type situations, provide positive empirical support for the existence of an objective extended present, even though it cannot be relativistically invariant, and so it favours Lorentz's theory over STR. More on this will follow a few pages below.

Appealing to the conventionality of the way distant clocks are synchronized in STR

It may appear to one, as, I confess, it appeared to me once, that the relativity of simultaneity rests on the method of synchronizing distant clocks suggested by Einstein, i.e. with light-signals, and that this is, as Einstein himself was ready to admit, just a convention, so it should not be treated as a metaphysical matter of fact, and so the relativity of simultaneity may perhaps be resisted on this ground. Indeed, this method of synchronizing distant clocks is standardly

[215] Cf. Popper 1982, Bell 1987, Tooley 1997, and Craig 2001..

called "the Einstein convention", and rightly so, for it is not made necessary by empirical facts. Now, one might argue, the relativity of simultaneity is equivalent to a hard metaphysical statement, i.e. the nonexistence of an objective present which could serve as the locus of becoming, and such a metaphysical statement should be objective, and not a matter of convention.

There is certainly something to this objection.

There is no physical experiment that could prove the invariance of the *one way* speed of light, when switching from one inertial frame of reference to another. The Michelson-Morley experiment, which textbooks usually cite as the empirical grounding for the invariance of the speed of light, shows only the isotropy of the *two way* speed of light, i.e. that the time it takes for a ray of light to travel from one end to the other of the arm of the interferometer, and then back again to the very point from where it started off, does not depend on the direction in which the arm of the interferometer is pointing. It is not a peculiarity of this particular experiment. It is easy to see that for measuring the one way speed of light one already has to know how to synchronize distant clocks, otherwise there would be no way of telling how much time it took for light to travel from one point to another. So one cannot argue for a method of synchronizing distant clocks on the basis of an empirical finding about the one way speed of light, since such an argument would be insuperably circular.

The Einstein convention supposes that it takes the same time for light to travel from A to B as from B to A. This supposition is very natural if we approach inertial frames of reference as Einstein did, i.e. if we assume that they are equivalent in the sense that they have an equal claim on capturing what we should mean by space and time. Suppose, however, that there is an Aether, a frame of reference at absolute rest. It is natural to suppose that the speed of light is isotropic relative to this frame of reference (since Aether was supposed to be the medium in which electromagnetic waves propagate). Now consider a frame that is moving at a constant speed relative to the resting frame. If the one way speed of a lightray doing a roundtrip was isotropic in the resting frame, then it could not be isotropic, unless it is measured by deformed measuring devices, in the moving frame.

The only thing we can know for sure is that if a lightray travels from point A to point B, hits a mirror in point B, and travels immediately back from B to A, and this whole process lasts from t_1 to

t_2, then the time when the lightray is reflected from the mirror in point B is $t_1+\varepsilon(t_2-t_1)$, where $\varepsilon \in [0,1]$[216]. Einstein proposed to set ε to .5, uniformly, for all inertial frames of reference. On Lorentzian terms, however, if $\varepsilon = .5$ in one frame of reference, then $\varepsilon \neq .5$ for all frames that are in motion relative to it. The impossibility of measuring the one way speed of light without assuming a method of synchronizing clocks entails that there is no empirical way of determining ε. ε can be set to any value between 0 and 1 without contradicting any empirical finding of physics.

If A and B are the two ends of an arm of an interferometer, for example, then ε determines which point of A's worldline is simultaneous with the point of B's worldline at which the lightray emitted form A reaches it and is reflected back. If there is no empirical way of determining ε, then there is no empirically knowable truth about which point of A's worldline is simultaneous with the lightray being reflected back by the mirror at B.

If there is no way of determining ε empirically, then, the choice of ε is, in a perfectly good sense, a matter of convention. Different ε's for different frames of reference can be set in such a way that the relativity of simultaneity disappears; the simultaneity planes of all inertial observers will be the same.

Now is it a good argument against the thesis of the nonexistence of an objective present? Surely, it isn't. It is true that the argument from STR to the nonexistence of an objective present that I have presented rests on the method of synchronizing distant clocks proposed by Einstein. And it is true that this method is not dictated by hard physical facts. Rather, this method is conventional.[217] But, quite obviously, it is already a problem for the advocate of an absolute present if the method of synchronizing clocks, that is, the simultaneity of spatially separate events, is a matter of convention. In fact, the problem gets deeper if one hopes to refute the argument from STR to the nonexistence of an open future on the ground that it is based on a conventional definition of simultaneity. One might do away with the disagreement about simultaneity between observers who are in motion relative to each other by pointing out that it rests on a convention that can be changed without any conflict with

---

[216] Reichenbach 1924, Definition 2 on p. 26.

[217] But see David Malament (1977) arguing that the choice of ε to be .5 is the only realistic candidate for a criterion of simultaneity if simultaneity is to be an equivalence relation defined within the inherent geometry of Minkowski spacetime.

empirical data, but in return, one would find that the simultaneity of spatially separate events is a matter of convention *even for one single observer*, so, as far as the prospects of presentism are concerned, one is out of the frying-pan into the fire.

Decomposing the spacetime of the general theory in the hope of obtaining objective time

The third possibility of resisting the conclusion derived from STR that there is no such thing as an objective extended present is to be found in the context of the general theory of relativity.

The difference between the special and the general theory is that the latter takes matter into account. A presentist may hope that the arrangement of matter breaks the symmetry of the observers of the flat spacetime of special relativity in a way that caters for his metaphysical purposes. Matter, in the tenseless, relativistic point of view, is a web of four-dimensional worldlines. If it is organized in some specific way, if there is an inherent directionality in the web of worldlines, then there is a unique natural foliation of spacetime into simultaneity planes. But, as Simon Saunders put it,

> the presentist will *literally* need a river for there to be time, according to his metaphysics.[218]

The first thing to note about the project of finding a physically unique foliation of spacetime, and taking its t parameter to be universal time, is that it may lead to an account of time whose generality is, in some sense, constrained. It may turn out to be true of our universe but may be false of others. Who cares about other universes?, one might ask.[219] Well, at the first glance, a presentist, or an A-theorist, in McTaggart's sense, seems to take the proposition that time flows to be a conceptual truth, which should be true not only in the actual world but in all worlds possible. McTaggart himself, at least, thought that it is a conceptual truth that time, if there is time, must flow. He came to the conclusion that time is unreal exactly because he thought it cannot flow. Now it seems that there are nomologically possible worlds in which spacetime surely cannot be foliated into spacelike hypersurfaces at all. Gödel found solutions to

[218] Saunders 2002.

[219] John Earman (1995), for one, has actually asked this question in response to Gödel.

the Einstein field equations of gravitation in which there are closed timelike curves,[220] which means that they may contain worldlines that bend back into their own past. Now, as Dennis Dieks noted in a recent article[221], it follows from this, that *a global linear flow* is not one of those "essential features" of time "which are determined by the Einstein equations only". Einstein's reaction to these weird Gödel-type universes was that "It will be interesting to weigh whether these are not to be excluded on physical grounds".[222] Indeed, geometrically well-behaved (globally hyperbolical) spacetimes can always be foliated with Cauchy surfaces.[223] In 1979 Roger Penrose hypothesized that all physically possible spacetimes are globally hyperbolical.[224] This hypothesis is, however, is not unanimously accepted by the physics community. *Our* spacetime seems to be globally hyperbolical. But should this satisfy a presentist? As Dieks summarizes the situation:

> Gödel's argument for the 'ideality' of time, as he puts it, relies on the idea that if time 'objectively lapses' (if there is objective becoming), this should be an *essential* property of time, instantiated in all possible worlds. The Gödel universes are then relevant as counterexamples. If the set of possible worlds is restricted so as to exclude Gödel universes, objective passage may regain its status as an essential attribute of time. As a limiting situation we could consider taking only our own universe as possible: then everything existing in our world would exist necessarily. The actual characteristics of time in our world would thus by definition also be essential. This seems a too drastic curtailment of the scope of physical theory and a trivialization of the distinction between the essential and the merely contingent, however. Even if we are convinced that there is actually only one universe and if we are strict

---

220 Gödel 1949b. Earman said (see the previous footnote) he tried benevolently to locate and reconstruct an argument in Gödel's work as to why the impossibility of a flowing time in the Gödel universe would be relevant to the question whether our time in the actual world was flowing, but he hasn't found one. I think it is clear that if the flow of time is a conceptual issue, then Gödel universes are relevant to the conceptual issue without such an argument.

221 Dieks 2006.

222 In Schilp 1949, p. 668. Cited by Dieks ibid.

223 Geroch 1970.

224 Penrose 1979.

> empiricists, it makes sense to conceptually distinguish between the merely contingent and the essential, on the basis of the properties of (a set of) models of our theories.[225]

I think Dieks is right in so interpreting the bearing of Gödel's results on presentism. Unless extravagant spacetimes are ruled out on a principle whose generality is comparable to that of the theory of relativity, a global flow, which seems to be a precondition for presentism, is not an *essential* feature of time. But I am not sure that it is a real worry for a presentist if it comes out that an objective global flow is a contingent feature of our time, which could be otherwise if our world was different. Dieks's concern, however, may be that this cannot be the case, because time flows either essentially, or it doesn't flow at all.

The empirical findings that led to the special theory of relativity allowed for two interpretations. One was that temporal properties are relational properties which are meaningful only relative to a frame of reference, and there is no deeper truth about time and space. Given that physically equivalent observers disagree about the present, the present cannot really have the ontological significance that presentism attributes to it. The other was that there is a deeper truth about space and time that singles out a reference frame that is at rest in absolute space, and whose time is the true time, even though inertial observers appear physically equivalent, and the time of the restframe is ontologically significant in the presentist's sense. We may prefer the first interpretation on verificationist grounds. The empirical situation is interpretable without postulating an absolute resting frame. We may prefer the second interpretation on the ground that the first would require us to consider temporal and spatial co-ordinates measured with instruments that we know are deformed as characteristic of proper time and space. General relativity and general relativistic cosmology seems relevant to the choice between these two

---

[225] Dieks, ibid, footnote 4. On the ground of Dieks's last sentence, it is hard to see why Gödel needed his exotic solutions to the Einstein equations to argue for the conclusion that a global flow was not essential to time (in Dieks's sense). The flat spacetime of special relativity is equally a model of the theory embodied in Einstein's equations, and thus is equally a counterexample to the conceptual thesis about the (global) flow of time, provided that STR is really inhospitable to that, as Gödel argues at other places. This problem has been noted by Steven Weinstein in his review in *The Philosophical Review* of Palle Yourgrau's 1999 book on Gödel and the ideality of time.

interpretations because it may be the case that global considerations show that the second of these two interpretations is correct, despite of the invariance of local observers. Dieks's, and, if his interpretation is correct, Gödel's, intuition might be that the choice between these two interpretations should apply to all possible worlds (obeying the equations of Einstein's general theory). Although Dieks doesn't say exactly why a global flow should not be a contingent feature of time, I suspect he thinks it would be weird to interpret the same phenomenon, i.e. the empirical invariance of local observers, in the Einsteinian way in some possible worlds, and in the Lorentzian way in others.

Although I think, in general, a presentist could be happy with a time that flows contingently, towards the principle that our choice between Einstein and Lorentz should be valid throughout all possible worlds I am sympathetic.

If this principle is accepted, however, then it must be admitted that the evidence in favour of Lorentz that the spacetimes that have a physically unique foliation provide is strong but nondecisive, whereas the evidence that the Gödel universes provide in favour of Einstein is decisive. If there is a unique foliation, then even a verificationist should take it as an indication that different inertial observers are not on par, after all, so from the fact that they disagree about the temporal classification of events it does not follow that the temporal classification of events is ontologically insignificant. It doesn't mean, however, that we can be sure that the spacelike hypersurfaces into which the unique foliation slices up spacetime represent the Newtonian absolute space at different successive instants of Newtonian absolute time, whose progression corresponds to an objective tide of coming to be. It only shows that the argument from the empirical invariance of local observers, and their disagreement about simultaneity, against the hypothesis that they do, is not a good argument. The Gödel universes, on the other hand, clearly exclude the hypothesis of a globally flowing absolute time, and there is nothing more to it.

But let us suppose, for the moment, that Gödel universes are excluded on the ground of a principle whose generality is comparable to that of relativity theory itself, and thus rightly restricts the models of the Einstein equations that should be considered as the spacetimes of possible worlds.

Then one might come up with the suggestion, sounding quite natural, that, in principle, the spacetimes of possible worlds all allow for a foliation into a successive series of spacelike hypersurfaces that are all orthogonal to the worldline of the centre of mass of the whole world.

If, however, we construe absolute time as the time of the centre of mass frame, then again, the generality of our account will inevitably be compromised, for it surely won't work for infinite worlds. A centre of mass an infinite universe is not very likely to have. But, unlike Gödel universes, which are now supposed to have been ruled out, infinite universes do not constitute a clear counterexample to the hypotheses that time is absolute and has a global flow, even though the preferred frame cannot be identified in them the way that has been suggested. The problem is that even in the case of finite universes one would already need to assume a definition of simultaneity to unambiguously define a centre of mass, so the definition of time as the succession of three-dimensional hypersurfaces orthogonal to the worldline of the centre of mass will be circular. Moreover, it may be the case that a finite universe has no centre of mass at all. Consider, for example, the standard expanding Friedmann-Robertson-Walker solution to the Einstein equations, underlying classic Big Bang cosmology. A snapshot of a Friedmann-Robertson-Walker universe (note that this already assumes a notion of simultaneity) can be imagined as the three-dimensional surface of a four-dimensional sphere, finite, but unbounded. Now assume, as it is actually assumed in the FRW model, that the distribution of matter in it is homogeneous and isotropic. Obviously, there is no centre to this mass-distribution, as there is no centre of mass of the homogeneously distributed matter of the surface of a ball, *within the surface*.

But even if there is no uniquely definable centre of mass, there may be a mean motion of matter, the "river" that Saunders mentioned. Time then would be defined as the succession of spacelike hypersurfaces that are orthogonal to the worldlines representing this mean motion.

This notion would, however, be dependent on the averaging method having been chosen. Dieks cites Gödel who already raised doubts whether the time so definable is a fit candidate for being the absolute time, because he thought the averaging process may contain "more or less arbitrary elements (such as, e.g., the size of the regions or the weight function to be used in the computation of the mean

motion of matter)", which he thought make it unlikely that a "precise definition which has so great merits, that there would be sufficient reason to consider exactly the time thus obtained as the true one".[226]

However, even if it comes out that the general theory fails to provide us with a method of identifying a preferred frame in all possible worlds, it is interesting to know whether it is identifiable in some of them, and it may be particularly interesting to know if there is a natural global time definable within the geometry of the spacetime of our actual universe, which may be a good candidate for being our absolute time. All the empirical evidence that, to some, seemed to have shown that there was no such time came from our world, after all. So if our world turns out to have a physically preferred frame, then this evidence is discredited.

Moreover, despite of what has been said about the generality of choice between the two interpretations of the empirical equivalence of inertial observers, it is not inconceivable at all, that we inhabit a universe which has a dynamic ontology contingently. So we have all reasons to explore whether it has a natural preferred frame.

Now there is cosmological reason to think that there is a natural way to define absolute time in our actual world. There is reason to think that, on the large scale, neglecting smaller scale inhomogeneities and unisotropies, the Friedmann-Robertson-Walker solution to the Einstein equations is an adequate representation of our spacetime. Now, assuming that there is a Big Bang-type singularity in the past of such universes, it is natural to define absolute time as a function of the radius of the expanding universe. The spacelike hypersurfaces, the succession of which represent the tide of becoming, then, are the successive spherical boundaries of an expanding four-dimensional ball, the one we considered a few paragraphs above. Time, therefore, can be defined on the basis of a property intrinsic to the spacetime geometry, i.e. its curvature.[227]

Arguably, the preferred reference frame is even experimentally identifiable.

One of the main empirical evidences that support Big Bang cosmology (alongside cosmic redshift), whose general relativistic basis

---

[226] Dieks ibid. Saunders (2002) expresses similar worries.

[227] A prominent advocate of so defining absolute time is Quentin Smith; see p. 140 of his 1989. Universal time, in fact, figures in the Robertson-Walker metric: $d\tau^2=dt^2+R(t)^2((dr^2/1-kr^2)+r^2d\Omega^2)$. If there was a Big Bang, its t parameter can be identified with the age of the universe.

was laid by Friedmann, is the cosmic microwave background radiation with a thermal 2.73 Kelvin black body spectrum, and an isotropy of 1 in 100,000, that was discovered in 1965 by Arno Penzias and Robert Wilson.[228] That background radiation was a prediction previously made on the ground of the Big Bang cosmological theory by several theoretical physicists.[229] The theory suggested that in the early universe electromagnetic radiation (photons) constantly interacted with charged matter (subatomic particles, i.e. electrons and protons), and formed a hot plasma in a thermic equilibrium, which, as the universe expanded, cooled until the formation of electrostatically neutral atomic matter (hydrogen) from electrons and protons became possible, as a result of which matter and radiation "decoupled", and the universe became transparent to radiation. This did not happen at a specific location, rather, it happened everywhere, at a certain time of the life of the universe (when it was about 380,000 years old), when the temperature of the universe dropped to approximately 3,000 Kelvin. The photons that were set free in the transition from plasma to atomic matter have filled out all space and have cooled ever since the decoupling happened, as has the whole universe, as a consequence of its expansion. What we are predicted to detect, then, is a radiation of much lower temperature (informative of the age and rate of expansion of the universe), which presents itself as coming isotropically from no particular source at all. And this is exactly what has been found.[230]

Now different observers certainly do not see the cosmic background radiation the same way. With respect to its spectrum, isotropy holds only in one frame of reference. In every other frame background radiation is redshifted in the direction of the movement of the frame relative to the previous one, and blueshifted in the opposite direction. Arguably, if Big Bang cosmology is correct, then it provides us with an experimental method of singling out a reference frame whose present is locally tangential to the spherical boundary of the expanding ball of our largely FRW-type universe, that is, whose

---

[228] For which they were later awarded the 1978 Nobel prize.

[229] Most prominently by Robert Dicke, George Gamow and Ralph Alpher from 1946.

[230] Although it is not my purpose here to recapitulate the empirical evidence in favour of Big Bang cosmology , it might be interesting to note that alternative cosmological models equally tried to account for the background radiation, but with lesser success. The steady state model of the universe, for example, lost its popularity in part because, although it succeeded to explain an isotropic background radiation of a comparable temperature, it failed to account for its perfect black body spectrum.

simultaneity planes are orthogonal to the worldline that locally corresponds to absolute rest, and represents the flow of absolute time.

Appealing to quantum non-locality—and what Schrödinger's cat has to do with it

In the last paragraph of a 1998 article about the prospects of tensed vs. tenseless theories of time John R. Lucas made the following concluding remark about quantum mechanics:

> It is too soon to suppose that quantum mechanics is the last word in physics, or that the way it is interpreted by me is the way it ought to be interpreted, but at least at the present time it looks as if a tensed view of time is in fact a view required not only by our ordinary untutored experience, but as a fundamental feature of the fabric of the physical universe.[231]

By contrast, Craig Callender, in his draft contribution to the yet to be published Craig-Smith anthology on absolute simultaneity writes this, quite sarcastically:

> Quantum mechanics seemingly offers something to everyone. Some find free will in quantum mechanics. Others discover consciousness and value. Still others locate the hand of God in the quantum wavefunction. It may come as no surprise, therefore, to hear that many believe quantum mechanics implies or at least makes the world more hospitable to the tensed theory of time. Quantum mechanics rescues the significance of the present moment, the mutability of the future and possibly even the whoosh of time's flow.…[T]he kind of reasoning underlying these claims is at least as desperate as that finding freedom, value, the mind and God in quantum mechanics—which is pretty desperate.[232]

To me it seems that Lucas overstates the support quantum mechanics gives to presentism, and Callender definitely underplays it.

---

[231] Lucas 1998, p. 43.
[232] Callender, forthcoming.

The argument from quantum mechanics to an objective global flow of time comes in two versions. The simpler version appeals to the quantum collapse *simpliciter* and claims that it brings objective becoming into the world. The more sophisticated version appeals to *the non-locality of the collapse* of the wave-function, or, if one is reluctant to take the collapse of the wave-function seriously, then more generally: the non-locality involved in the correlatedness of the states of spacelike separated pairs of quantum objects with a common past, like those in the Einstein-Podolsky-Rosen thought experiment.

I will argue that the simpler version of the argument is flawed, but the more sophisticated version of the argument may prevail on some interpretations of quantum mechanics that are seriously considered presently by both physicists and philosophers of physics, even on interpretations that do not have the collapse of the wave-function. Not on all, unfortunately.

Lucas is a prominent current advocate of both the simpler and the more sophisticated versions of the argument. Let us consider the simpler version first.

The simpler version

Reflecting on the argument from the special relativistic relativity of simultaneity to the unreality of objective becoming he writes this:

> The Special Theory is not the last word in physics, and its Principle of Equivalence does not have to hold universally, and does not rule out any preferred hyperplane of simultaneity. *In fact, other physical theories rule it in*. Most cosmologists use a version of the General Theory with boundary conditions that determine a universe-wide world time. Admittedly, cosmological theories are speculative, and liable to change radically: but the mere fact that cosmologists at present postulate a world time is enough to discredit any argument from the Special Theory that there is something unscientific in a world-wide hyperplane of present simultaneity.

We have already been through this. But he goes on, and here comes the simple version of the argument from quantum mechanics:

> But physics goes further. It not only defeats the would-be defeaters of the tense theory, but offers positive support. *Quantum mechanics*, if it is to be interpreted realistically, *distinguishes a probabilistic future of superimposed eigen-states from a definite past in which each dynamical variable is in one definite eigen-state, with the present being the moment at which*—to change the metaphor—*the indeterminate ripple of multitudinous wave-functions collapses into a single definite wave.* Admittedly, many of those who think about quantum mechanics are not realists, and admittedly again, there are horrendous difficulties in the way of giving a coherent account of the collapse of the wave-function. But an obstinate realism, as well as a slight sympathy for our feline friends, precludes my envisaging any long period in which Schrödinger's cat could be half-dead and half-alive, and this whether she be in a laboratory in Europe or on some planet circling Betelgeuse. *There is a definite fact of the matter*, there as much as here, *whether or not we are dealing with a superposition of functions or one definite eigen-function. And hence there is a unique hyperplane advancing throughout the whole universe of collapse into eigen-ness.*[233]

Lucas here assumes that the collapse of the quantum mechanical wave-function (or state-vector) is to be interpreted realistically, but it is not necessary to challenge this assumption to see that what he says is untenable.

The quoted passage, especially the italicized parts of it, strongly invite a reading which is really wild, and which, I believe, cannot be what Lucas means. Yet it may be useful going through it quickly to do away with possible misunderstandings.

From the last two sentences of the paragraph it might seem that Lucas thinks that the "unique hyperplane advancing throughout the whole universe of collapse into eigen-ness", i.e. the objective universal present, consists exclusively of quantum collapses. The second sentence of the paragraph suggests that Lucas thinks that on one side of this hyperplane there are only superposition states, this is the future, and on the other side, which is the past, there are only eigenstates. As if there was a one-to-one correspondence between the ontological openness of future events and superposition states, on the one side, and between the ontological fixity of past events and "eigen-

[233] Ibid. Emphasis mine.

ness", on the other. (Indeed, Lucas seems to coin this term as a synonym of ontological fixity.)[234] Becoming and quantum collapse, on this picture, would be two names for the same thing.

This view would be absurd. I don't think that this is what Lucas is in fact saying, but it took me quite a while to get clear on his position. The simple identification of the three ontological stages that presentists associate with the three tenses with "superposition-ness", collapse, and "eigen-ness", that this paragraph seems to invoke, might seem attractive to some. In explaining why it would be absurd, I assume some familiarity with the formal scheme of quantum mechanics and its interpretation by John von Neumann.[235]

Lucas is a realist about an interpretation of quantum mechanics which is not very different from the original von Neumann interpretation. Like many, Lucas is committed to the view that the collapse of the wave-function is not exclusively triggered by the performance of measurements, but happens continuously and spontaneously.[236] Otherwise his interpretation is the same.

Now, if Lucas's position was the one I outlined above, then his realism about the wave-function would be very selective. To start with, he would have to be committed to the view that reality is incompatible with "superposition-ness".

Reading his remark about Schrödinger's famous cat, one might have the impression that this is really what he thinks. He says, *"an obstinate realism, as well as a slight sympathy for our feline friends, precludes my envisaging any long period in which Schrödinger's cat could be half-dead and half-alive"*.

The reference to the length of the period in which the cat could still be in a half-dead and half-alive superposition state, however, should discourage us from so interpreting his position. If Lucas thought that the triad of "superposition-ness", collapse, and "eigen-ness" could be mapped onto the triad of ontological openness, fixity just being acquired, and unchangeable ontological fixity, and thus on that of futurity, presentness, and pastness, then he would have to hold that the cat can never *really* be in such a superposition state, *for any period of whatever length*.

---

234 A counterpart term could be "superposition-ness".

235 It is reviewed briefly in the Appendix.

236 So he is a spontaneous collapse theorist. About this see §11.7 of his new book, *Reason and Reality*, titled "The 'Measurement Problem'" (2006).

I think what he thinks is rather that reality and "superposition-ness" are not strictly incompatible, but are in some tension, which tension must be relieved, and is relieved, in all cases, within a short time after it has been built up. This reading is confirmed by the relevant chapter of his 2006 book.[237]

But this latter reading of what he says here is of course incompatible with the raw identification of becoming with the collapse of the wave-function, which the key sentences of the quoted passage seem to suggest.

If the "becoming is collapse" thesis was Lucas's position then he would be in trouble for two very obvious reasons.

One is that quantum mechanics seems to know about a smooth evolution of the wave-function between any two collapses, von Neumann's process 2. Lucas, on this reading of the quoted paragraph, would be a realist only about process 1, the collapse of the wave-function. Process 2 he would have to place in the future, which, as a presentist, he thinks does not exist. This would be a very unusual way of being a realist about quantum mechanics. There are interpretations of quantum mechanics which are selectively realist about the dual dynamics posited by von Neumann, but they are realist about process 2, the evolution governed by the linear dynamical law, and get rid somehow of process 1, the collapse. The "becoming is collapse" thesis would require one to think something like that process 2 has only ideal existence, that it exists only in the minds of physicists contemplating about the probabilities of the possible outcomes of future observations, which latter, when they will actually be performed, will always have definite results, corresponding to eigenstates, not superposition states. But it cannot be right. If the calculations of probabilities of the possible outcomes of a measurement to be performed on a system (on a multitude of identically prepared systems) at a future time prove correct empirically, then there must be an explanation for it. The most natural explanation for it is that, in the period that spans between the system's being prepared and its being measured, it really evolved in accordance with the dynamical law. If it was the case that the superposition state–collapse–eigenstate triad maps onto the future–present–past triad, then, quite absurdly, we would have to expect the wave-function to collapse in every single moment, as time advances,

---

[237] Ibid. §11.6.

rather than evolving smoothly, between its preparation and its measurement, and then the predictions based on the supposedly smooth evolution guided by the dynamical law would be useless. Or, alternatively, one would have to think that becoming is not happening continuously but only at times when the wave-function collapses, but this view would not sit well with presentism.

I am sure Lucas does not mean either of these. In §11.7 of his *Reason and Reality*, he says he thinks that, except in carefully designed laboratory experiments, quantum systems become "entangled" with each other, and "often 'collapse' into an *eigen*-state of the interfering system".[238] How often, Lucas doesn't say. But even if "very often", he must be a realist about the short periods between any two collapses, that is, he must be a realist about von Neumann's process 2. But in process 2 the quantum mechanical system evolves smoothly through superposition states. But then it is not true that a superposition state can exist only in the indefiniteness and not-yet-real-ness of the future, with the definiteness of the past corresponding to eigenstates only, and with the present, dividing the realms of indefiniteness and definiteness, being the exclusive locus of the collapse. But if it is not true, then becoming is not identifiable with the collapse of the wave-function.

The other problem with the view in question is that "eigen-ness" and "superposition-ness" are not properties that the states of quantum mechanical systems have intrinsically. These have meanings only relative to a physical property, or the operator representing it. Most operators representing different physical properties have different sets of eigenstates. This is behind Heisenberg's uncertainty principle. If a system has just been measured for a property, then, immediately after the measurement, the system is in an eigenstate of that property, and we know the value of that property in that state without ambiguity. We could also know the value of the other property we would like to know in the same state unambiguously only if this state was an eigenstate also of that property. But generally this is not the case. A state which is an eigenstate from the perspective of one property is often a superposition state from the perspective of another. So the raw identification of ontological openness (objective futurity) with "superposition-ness", and ontological fixity (objective pastness) with "eigen-ness", which Lucas seems to advocate on a

[238] p. 329.

superficial reading of his text, is untenable for this very simple but deep reason, too.[239]

I think the way Lucas formulates his view comes so dangerously close to this obviously fallacious position because this position is so obviously fallacious that Lucas does not feel it necessary to spend time on clearly demarcating his position from it.

His position, I believe, is subtler than the one we were discussing this far. I think what he means is not that becoming is identical with the collapse. I think every moment of the smooth evolution of the wave-function qualifies as an instance of objective becoming for Lucas. So I think he is totally happy with cases of objective becoming when we are dealing with superposition states equally before and after, and nothing discontinuous happens in the present. I think what he really means is only that the case when a collapse happens is special, because in such a case there is no doubt that an ontological change has taken place. Unlike other cases of objective becoming, quantum collapses are obvious marks of ontological transition. And, given that on the interpretation of quantum mechanics preferred by Lucas the collapse of the wave function is instantaneous, the instantaneous collapse of the wave function of extended systems picks out the preferred foliation of spacetime that corresponds to true time, the *"unique hyperplane advancing throughout the whole universe of collapse into eigen-ness"*. I think this is what he says.

Why would wave-function collapses be obvious marks of ontological transition?

On p. 332 of *Reason and Reality* Lucas writes:

> If quantum-mechanical systems are continually being confronted by a moment of truth, when various possibilities are winnowed out, leaving one definite state of affairs, there is an ontological difference between the future and the present and past.

I take it that it means that a collapse of the wave-function marks an ontological change, because otherwise it could not be the case that,

---

[239] This problem is mentioned also by Callender criticizing Lucas. This problem in itself could, in principle, be surmounted if an argument was offered for preferring the decomposition of the Hilbert space in terms of the eigenstates of a certain physical property. The "preferred basis problem" will resurface in the context of some interpretations of quantum mechanics discussed in the Appendix.

before the collapse takes place, a plurality of outcomes is objectively possible.[240] The quantum mechanical account of physical reality, the argument goes, could not be objectively probabilistic, if the future was not objectively modal. The transition from the multiplicity of possibilities to the uniqueness of actuality in the collapse is marked by the difference between the "superposition-ness" of the pre-collapse wave-function, and the "eigen-ness" of the post-collapse one, even if these qualities of the wave-function are meaningful only relative to a physical property. So the ontological status of future events must be different from those of present and past events exactly the way presentists say it is different.

Well, first of all, there is no consensus about the objectivity of the probabilistic nature of the evolution of quantum reality. There is a practically unanimous consensus on the impossibility of any local hidden variable theory, which, when supplemented, could make quantum mechanics deterministic, but non-local hidden variable theories (such as Bohmian pilot wave theories[241]) are still in the competition, and there are also other deterministic interpretations on which indeterminism is present only at the subjective level, as it is discussed in the Appendix.

But even if there was a consensus favouring, say, the classical von Neumann interpretation of quantum mechanics, that is, even if quantum mechanics was patently objectively and not just epistemically probabilistic, Lucas would be wrong. For quantum mechanics is a causal theory, and what he says confuses the ontological issue with the causal one.

If the evolution of physical reality is objectively, not just epistemically, probabilistic, it means that there are no hidden variables which are epistemically inaccessible, yet guide the evolution of quantum reality deterministically. So quantum reality would then be objectively indeterministic. But this concerns only the causal organization of physical reality, and bears no consequences on the ontological fixity/openness of the future. A block universe can have a causal organization, and it can be both deterministic and indeterministic. The ontological fixity of a future event in a block universe is consistent with its being a causal dangler.[242] If there are no

---

[240] The same view was expressed earlier by Karl Popper, cf. Popper 1982.

[241] Bohm 1957.

[242] And vice versa. Presentism is compatible with determinism: the unreality of a future event is compatible with its being determined by what has already become real.

hidden variables, if quantum mechanics is complete, it means only that there are objective causal danglers. For this, the set of all events the history of the universe consists of, future and past, can be ontologically homogeneous.

So I think we have to conclude that the simpler version of the argument from quantum mechanics to presentism fails on either reading of it. Now let us turn to the more sophisticated version.

The more sophisticated version

Ironically, the more sophisticated quantum mechanical argument purported to show that the Einsteinian relativity of simultaneity is in fact false, and so there may be an objectively open global future, grew out of a thought experiment which Einstein himself devised, in the hope that he can show that quantum mechanics is incomplete, and so reality may be deterministic.

He asked us to consider two particles with a common past, but now separated. Their common past is important to secure that their quantum states be correlated; for the sake of quantum mechanical description they count as two constituents of the same system. Now, according to Heisenberg's uncertainty principle, neither of the two particles can have a definite position and a definite momentum at the same time. But there is no objection against measuring the position of one of them, and the momentum of the other. However, given the conservation of momentum, applicable to the system consisting of the two particles, once the momentum of one is measured, the momentum of the other is known, too. Einstein thought it proves that even if it is true that the position and the momentum of the same object cannot be measured at the same time, it can *have* both a definite position and a definite momentum, contrary to Heisenberg's principle. So quantum mechanical description is incomplete, for there is an element of reality it doesn't account for, i.e. a definite value of the momentum while the position is definite. If quantum mechanics is incomplete, then it may be just a part of a fuller picture, which can be deterministic.[243]

The argument for absolute simultaneity arises from the fact that Einstein's explanation for the thought experiment, i.e. the incompleteness of quantum mechanics, has an alternative: the non-

[243] Einstein, Podolsky, Rosen, 1935.

locality of quantum mechanics. According to the incompleteness explanation the momentum of the particle whose position was measured had already a definite value before the momentum of the other particle was measured. According to the non-locality explanation, however, the first particle acquires a definite momentum only when the momentum of the second particle is being measured. The momenta of the two particles acquiring definite values, one as a consequence of the measurement, the other as a consequence of the first having acquired a definite value, are two spacelike separated events. Now, if the two particles are spacelike separated, these two events cannot be connected by any means of information transmission that respects locality (i.e. the principle that any effect should propagate from one point to another through the space between them, with the states of affairs that obtain at a given location affecting directly only their immediate neighbourhood), since no effect propagates faster than light. Yet, they are correlated. The non-local quantum mechanical explanation for their correlatedness must be essentially holistic. The two spacelike separated particles should be considered as a spatially extended whole, described with one single wave-function. The standard quantum mechanical explanation of the correlation would be that the measurement performed on one of the two particles is in fact a measurement performed on the whole object, and collapses the whole wave-function instantaneously. Now, given the spatial extendedness of the system it describes, the notion of 'collapsing the whole wave-function instantaneously' invokes absolute simultaneity.

Later the discussion concentrated on a physical scenario which brings out exactly the same theoretical problem as the one described above, but, rather than the momentum, concerns the spin of two electrons that were emitted from a common source in opposite directions. The spins of the two electrons can be measured with Stern-Gerlach magnets.[244] Once we know the spin of one of the electrons, we know the spin of the other without uncertainty, because of their correlatedness. In 1964 John Bell pointed out that any hidden variable account of the correlatedness of the spins of the two electrons, i.e. any explanation that assumes that quantum mechanics is incomplete and there is an underlying mechanism which is responsible for the correlatedness of the two spins, *which is assumed to*

[244] Bohm and Aharonov 1957.

operate respecting locality, leads to certain inequalities (the Bell inequalities) that are violated by the statistical predictions of quantum mechanics.[245] The empirical tests of these Bell inequality-violating quantum mechanical statistical predictions (which Bell and Abner Shimony dubbed "experimental metaphysics") have actually been performed, first in 1981[246], and now there is a wide consensus on the matter that there is no local hidden variable explanation for EPR-type phenomena. Whether or not quantum mechanics is complete, reality seems to be non-local.

This is why Karl Popper thought the empirical underdeterminatedness of the choice between Einstein's special theory of relativity and Lorentz's alternative interpretation was finally broken, and wrote in 1982:

> It is only now, in the light of the new experiments stemming from Bell's work, that the suggestion of replacing Einstein's interpretation by Lorentz's can be made. If there is action at a distance, then there is something like absolute space. If we now have theoretical reasons from quantum theory for introducing absolute simultaneity, then we would have to go back to Lorentz's interpretation.[247]

However, buying uncritically the explanation for the correlation in terms of the holism of the wave-function and its instantaneous collapse would be totally innocent of a fundamental conceptual problem of quantum mechanics, to the solution of which different "interpretations" have been developed. These interpretations of

---

[245] Bell 1964 and 1966. The locality condition that Bell actually used was analysed by Abner Shimony (1986) as the conjunction of the condition of parameter independence and that of outcome independence. The former means that the arrangement of the measuring apparatus in one wing of the experiment does not effect the outcome in the other wing. The latter means that the outcomes on the two sides are not probabilistically dependent on each other, assuming that all relevant information has been taken into account in the description. Violation of the Bell inequalities imply the violation of either of the two or both. The instantaneous collapse of the wave-function of the pair explains the correlation between them by the holism of the wave-function violates outcome independence. From the no-collapse theories discussed in the Appendix, the one that Bell considered, that is, Bohmian mechanics, violates parameter independence. (Cf. Bacciagaluppi 2001.)

[246] Cf. Aspect, Grangier, Roger 1981, and Weihs, Jennewin, Simon, Weinfurter, Zeilinger 1998.

[247] Popper 1982, p. 30. Quoted by Callender, ibid.

quantum mechanics differ in their realism or non-realism about the wave-function and its collapse, and on some plausible interpretations the instantaneous collapse of the state of systems with spacelike separated parts, which is the alleged basis for the vindication of absolute simultaneity, simply does not obtain. Some of the proponents of these interpretations (most famously Hugh Everett) claimed that their interpretation can accommodate EPR-like phenomena retaining locality (although not by explaining the correlation with a hidden mechanism respecting Bell's locality condition).

So to have a position on the question whether there is really an argument from EPR-like phenomena to absolute simultaneity one has to have a position of how quantum mechanics should be interpreted.

The measurement problem and the solutions to it favoured by significant portions of the scientific and philosophical community, that is, the main interpretations of quantum mechanics, are reviewed in the Appendix.

I myself haven't yet clarified my position on how quantum mechanics should be interpreted. In a parallel Everettian world I may currently be pursuing research goals directly concerned with the interpretation of quantum mechanics, but as far as this world is concerned, at least for now, I largely regard it as a question on which I need instruction from specialists.

Consequently, I do not have a finalized position on the question whether quantum mechanics supports absolute simultaneity and so contradicts the special relativistic symmetry of local observers. My impression of the state of the discussion is that it should be considered as an open question. I think the advocates of absolute simultaneity—contrary the claim that I have earlier quoted from Callender mocking the advocates of an A-theory of time who "desperately" appeal to quantum mechanics—may indeed hope that the matter will one day be decided their way. Yet, Popper's optimism that "the suggestion of replacing Einstein's interpretation by Lorentz's can be made" on the ground of the confirmation by Aspect *et al.* of the correlatedness of spacelike separated measurements, or Lucas's that quantum mechanics "rules in" absolute time, is premature, to say the least.

From the solutions to the measurement problem reviewed in the Appendix only some versions of the Everettian approach seem to be manifestly reconcilable with the equivalence of inertial observers (the

relativity of simultaneity).[248] The Everettian approach is, arguably, the wildest of all.

If this is so, I mean, if from the pool of interpretations that are endorsed by significant portions of the scientific and philosophical community only the Everettian interpretation is uncontroversially relativistically invariant (or if the Everettian approach makes quantum mechanics relativistically invariant in a uniquely appealing way, not diminishing the theories credibility), then this feature can be viewed

---

[248] There is always, of course, the option of declining from giving an interpretation to quantum mechanics in quite the sense the attempts discussed in the Appendix do. Among the founding fathers of quantum mechanics Niels Bohr was perhaps the most modest in this respect. The term 'Copenhagen interpretation' is vague, but it is most appropriately used to refer to his views. (Sometimes the term is used merely to refer to the view that quantum mechanical indeterminacy is objective. Sometimes it is used to refer undifferentiatedly to the most influential early ways of understanding quantum mechanics, predominantly those of Bohr, Heisenberg, Born, and von Neumann, which, however, do not fully fit together in a coherent way. Sometimes the view that the presence of consciousness collapses the wave-function is presented under the label of the Copenhagen interpretation.) Bohr's view, inspired by Kant's *Critique of Pure Reason*, was that quantum mechanical weirdness, that is, violation of very intuitive principles that solidified during the two and a half centuries of classical physics, mainly arises from the fact that our *a priori* weaponry to form concepts, which predetermines what shape our understanding of the objective world can take, manifested in the concepts of classical physics, is simply inappropriate to deal with the reality grasped by the quantum mechanical calculus. (This thought, involving the conviction that we have only the classical concepts to formulate unambiguous thoughts of reality, led Bohr to his two famous methodological principles, complementarity and commensurability, the former meaning roughly that only a duality of mutually exclusive descriptions using classical concepts can grasp quantum mechanical reality, the latter meaning that quantum mechanics should turn out to be a generalization of classical physics, in the sense that in cases when the effect of the inappropriateness of classical concepts is not very significant—in the case of "large quantum numbers"—quantum mechanical predictions should approximate classical ones.) If so, then there is no point even in asking whether we should be realist about the elements of the calculus, for example, whether we should be realist enough about the non-local collapse of the wave-function to think that it might undermine the relativity of simultaneity. Of course, we shouldn't. Then, even without being a positivist, one is forced to adopt a rather positivist attitude towards quantum mechanics. Due to our inability to form concepts about the quantum world the adequate way, we can only view the quantum mechanical calculus as a mere statistical book-keeping device for the predictable outcomes of measurements to be performed on multitudes of identically prepared systems, without the hope of deriving a metaphysics from it. (The occurrence of paradoxes, conceptual conflicts shouldn't take us by surprise.) As far as I can tell, the consistent histories approach to quantum mechanics, standing alone, not embedded in an Everettian interpretation, as it is in the case of, e.g., Saunders's theory, can be viewed, in some respect, as a *renaissance* of the metaphysical modesty of the Copenhagen interpretation (understood as referring to Bohr's approach).

as an argument for preferring it to its alternatives, despite of its excessive metaphysics. In this spirit writes Saunders the following:

> Is the Everett approach believable? But I know of no other that is. We can do no better than seek a coherent and systematic interpretation of physical theory. It should respect our pre-theoretical opinions that we are firmly attached to; and it should respect hard-won theoretical principles as well. ... [A]mong our theoretical principles I prize the relativity principle, and the principles of quantum mechanics. Everett's ideas hold out a radical and austere way of combining them; that is the reason to pursue them.[249]

But what if we consider the relativity of simultaneity not as a "hard-won theoretical principle" but as a principle that concurs with Lorentz's, and ourselves as being in the situation of making a choice between the two?

As it was discussed above, the Lorentz-Fitzgerald theory is empirically equivalent to the special theory of relativity. There is no empirical reason to prefer one to the other, therefore, non-empirical principles of theory-choice need to be invoked in order to justify a choice between them. The principle that supports STR is related to its simplicity. The phenomena that the two theories account for equally well has a symmetry, the phenomenological equivalence of inertial observers. Both Lorentz and Einstein account for this symmetry, but Einstein does it in a simpler way, with a theory that is itself symmetric at the metaphysical level. Lorentz's theory is asymmetric, and the asymmetry is due to the postulation of extra metaphysical entities—absolute space and time. There is one among the infinitely many inertial frames of reference, whose co-ordinates describe absolute space and time. The theory then explains why this metaphysical asymmetry does not give rise to phenomenological asymmetry. By this very explanation—this is the other side of the same coin—the theory explains why these postulated metaphysical entities are empirically unverifiable. The simplicity-oriented reason to favour STR can now be stated both in an Occamist and in a verificationist language: the difference between the two empirically equivalent

---

[249] 2000, p. 11. (As it is the case with all papers of Saunders cited in this thesis, the page number refers to the pdf-version of the article that appears on Saunders's homepage.)

theories is that one of them postulates metaphysical entities without which the other accounts for same the phenomena equally well, and precisely because the phenomena can be accounted for without the postulation of these entities, the existence of these extra entities are empirically unverifiable.

This much would be enough to prefer Einstein to Lorentz, had the differences in the metaphysics of the two theories been indifferent with regard to the two theories' compatibility with other theories we value highly.

One such theory to consider is quantum mechanics. If quantum mechanics comes in different versions, and we have to choose from them, then it is not a good argument for a version of quantum mechanics which postulates extra metaphysical entities (like parallel worlds that are created continually in billions) that only this version can be squared with Einstein, if there are other versions that do not postulate these extra entities and can be squared with Lorentz, if our previous preference for Einstein rather than Lorentz was based solely on the simplicity principle.[250]

---

[250] The Everettian parallel universes *prima facie* seem to be on a par with the absolute space and time of Lorentz's theory. There are interpretations of quantum mechanics that do without them, and their existence seems empirically unverifiable, since parallel Everettian universes do not interact, by hypothesis (no decoherent branch of the universal wave function has any effect on any other). Yet, Max Tegmark proposed an experiment to verify their existence, known as "the quantum suicide experiment" (Tegmark 1998). This is how he describes the experiment: "The apparatus is a 'quantum gun' which each time its trigger is pulled measures the z-spin of a particle. It is connected to a machine gun that fires a single bullet if the result is 'down' and merely makes an audible click if the result is 'up'. ... The experimenter first places a sand bag in front of the gun and tells her assistant to pull the trigger ten times. All [interpretations of quantum mechanics] predict that she will hear a seemingly random sequence of shots and duds such as 'bang-click-bang-bang-bang-click-click-bang-click-click'. She now instructs her assistant to pull the trigger ten more times and places her head in front of the barrel. This time the 'shut-up-and-calculate' [the Copenhagen interpretation] have no meaning for an observer in the dead state...and the [interpretations] will differ in their predictions. In interpretations where there is an explicit non-unitary collapse, she will be either dead or alive after the first trigger event, so she should expect to perceive perhaps a click or two (if she is moderately lucky), then 'game over', nothing at all. In the MWI [Everettian multiple worlds interpretation], on the other hand, the...prediction is that [the experimenter] will hear 'click' with 100% certainty. When her assistant has completed this unenviable assignment, she will have heard ten clicks, and concluded that the collapse interpretations of quantum mechanics are ruled out to a confidence level of $1-0.5^n \sim 99.9\%$. If she wants to rule them out 'ten sigma', she need merely increase n by continuing the experiment a while longer. Occasionally, to verify that the apparatus is working, she can move her head away from the gun and suddenly hear it going off intermittently." I think it is uncontroversial that the quantum suicide experiment is

Generally, if the simplicity principle has to have purchase here, and in similar cases, it has to be applied to empirically equivalent *coherent combinations* of theories. If the simplicity principle is to favour STR, then STR in combination with the simplest version of quantum mechanics that is compatible with it has to beat, in terms of simplicity, Lorentz's theory in combination with the simplest version of quantum mechanics that requires a preferred frame of reference.[251]

At the moment, I have no position on the question which combination wins in such a comparison. My superficial impression is that Lorentz's theory in combination with Ghirardi, Rimini and Weber's spontaneous collapse theory[252] (of which no sufficiently general relativistically invariant version has been produced yet) is arguably simpler overall than STR combined with the Everett interpretation[253] (either in Albert and Loewer's[254] or in Saunders's version[255], even if I must admit that STR combined with Saunders's version of the Everettian interpretation is much superior aesthetically), but I do not want to press this point.[256] What I only

---

capable of verifying the Everettian interpretation. It has, of course, never been performed. The reporter of New Scientist (issue 2113, 20 December 2007, p. 50, available on-line from Tegmark's website) remarks that Tegmark suggested to him that "Perhaps I'll do the experiment—when I'm old and crazy". Even if he does so, however, very probably it won't be very instructive for us. Even if MWI is correct, 99.9% of our selves will read in the newspapers that he died in trying to verify his beloved theory. He will be the only one who will not have the vast majority of his selves believing that the experiment ended tragically—for the simple reason that those copies of him will be no more.

[251] It should be noted that the relevant combinations are not those of Lorentz's theory with interpretations of quantum mechanics that *require* absolute simultaneity, but with those that *tolerate* it. But, of course, since if a combination involving Lorentz's theory is to beat a combination involving STR, it should be because the version of quantum mechanics that is involved in the former combination is simpler than the version that is involved in the latter, and since any interpretation of quantum mechanics that does not require, only tolerates, absolute simultaneity can be combined with both Lorentz's and Einstein's theory, it is enough to consider the combinations of Lorentz's theory with interpretations of quantum mechanics that require absolute simultaneity.

[252] Ghirardi, Rimini and Weber 1986.

[253] Everett 1957.

[254] Albert and Loewer 1988, Albert 1992.

[255] Simon Saunders 1995, 1996, 1998, 2000.

[256] Some might respond to this that STR in combination with the consistent histories approach (cf. Omnès 1994), not embedded in an Everettian interpretation but standing alone, is simpler than Lorentz in combination with GRW. This, I think, is quite uncontroversial. However, if the consistent histories approach in itself is to solve the measurement problem, then this work is left fully to decoherence, and this work, I think, decoherence alone cannot perform. (That was the reason why the consistent histories

wanted to emphasize is that if the metaphysical claim concerning the non-existence of absolute time is to be based on a theory that is preferred to a rival, empirically equivalent theory, which has absolute time, on the basis of non-empirical principles of theory-choice, then those principles should be applied not to separate theories, but to coherent combinations of theories in order to justify such metaphysical commitments.

Finally, I would like to add one point, not related to quantum mechanics, to the question of the compatibility of the two empirically equivalent theories, Einstein's and Lorentz's, with other theories that we value. If it was the case (i) that the special theory of relativity is incompatible with objective becoming, (ii) that objective becoming is a precondition for libertarian freedom, and (iii) that libertarian freedom is a precondition for rationality, then we would have a much stronger reason to favour Lorentz's theory than the simplicity-based reason we have to favour Einstein's, for, I believe, the theory that our theories are the results of rational reflection we value really highly. For (iii) I have argued extensively in chapter 5. If either (i) or (ii) is false, then STR is no threat to libertarian freedom.

How to cope with the relativity of simultaneity – local presentism

In the preceding sections we were engaged in the business of trying to resist the relativity of simultaneity on behalf of presentism—by appealing to quantum mechanical phenomena that, some claim, are best accounted for by adopting a theory that contradicts the equivalence of local inertial observers, or, failing that, accepting the symmetry of local observers and trying to obtain absolute time globally, cosmologically, or, failing that too, simply by claiming that the special theory of relativity has an empirically equivalent alternative, Lorentz's theory, and that there is nothing that would literally force us to choose the former and the relativity of simultaneity with it.

Some say it was unnecessary. If the philosophical effort to resist the relativity of simultaneity was motivated by the fear that accepting it objective becoming would be lost, then, these philosophers argue, it was a false alarm, because the relativity of simultaneity is compatible with objective becoming.

---

approach in itself is not discussed among the interpretations of quantum mechanics in its own right in the Appendix.)

Dennis Dieks on the mutual irrelevance of the relativity of simultaneity and our experience of the passage of time and becoming

Perhaps the strongest motivation to maintain objective becoming is that we seem to experience, as Dennis Dieks puts it, "that history unfolds and events come to being".[257] It is hard to believe that this is illusory. The argument that is supposed to force us to contend that it is illusory was summarized by Gödel (already quoted) as follows:

> The existence of an objective lapse of time…means (or, at least, is equivalent to the fact) that reality consists of an infinity of layers of "now" which come into existence successively. But, if simultaneity is something relative in the sense just explained, reality cannot be split up into such layers in an objectively determined way. Each observer has his own set of "nows", and none of these various systems of layers can claim the prerogative of representing the objective lapse of time.[258]

Central to this argument is the lack of an objective global temporal ordering of events. However, on a closer look, Dieks argues, the experience that "history unfolds and events come to being", does not involve, nor is dependent on, a global temporal ordering of events. Our idea of becoming is derived from a purely *local* experience of temporal ordering, while the special theory of relativity says only of the *global* temporal ordering of events that it is relative to the choice of a frame of reference. A global temporal ordering of events must rest on the notion of simultaneity, and simultaneity is frame-relative. But simultaneity, Dieks says, plays no role in our experience of the passage of time.

Dieks's "epistemological critique of the relevance of simultaneity" draws on two assumptions. One is that there is no action at a distance. The other is that perception is a local process.

The first of these two assumptions is accepted by everyone accepting STR. The second perhaps may be challenged.

Perhaps observations, rather than being exactly local, supervene on events in a small but extended spacetime region, whose

---

[257] Dieks 2006.

[258] Gödel 1949a, p. 557.

dimensions are determined by that of the body. Even so it is a question whether simultaneity within this small spacetime region makes any difference to the observation. Dieks says it is quite implausible to suppose that it does, and even if it does, representing observations as point-events is a very good approximation.

Now if our experience is temporally coarse-grained, as it is according to the psychological thesis of the 'specious present'[259], then Dieks is very probably right: this coarse-graining makes the ambiguities in the temporal ordering of experiences possibly caused by the extendedness of the body and the relativity of simultaneity irrelevant.

Now the lightcone structure is relativistically invariant. Events in the upper (future) lightcone, and events in the lower (past) lightcone are unambiguously ordered temporally with respect to the apex. Dieks emphasizes that experientially this is reflected in the fact that in the apex we could have veridical memories of any event in the lower lightcone, and in any point in the upper lightcone the apex event could be veridically remembered. There is a question only about the temporal order of the events outside both cones, relative to us in the apex. This question is of course the question of simultaneity. Einstein said that this question should be answered stipulatively, and the stipulative definition he gives to simultaneity makes simultaneity frame-dependent. Dieks argues that this question can be left unanswered.

When Einstein suggests that the question of the temporal ordering of distant events relative to the event here and now should be decided by synchronizing distant clocks to our clock here with light-signals, he takes the one-way speed of light to be invariant. As it was discussed earlier, there is empirical evidence only of the invariance of the two-way speed of light. Taking the one-way speed of light to be invariant too is equivalent to taking Reichenbach's ε to be ½.[260] But of course if no effect propagates faster than light, and if observation is a local process, then our experience is invariant under different choices of the value of ε. As far as our experience is concerned, events spacelike separated from us can be left unordered temporally.

If so, Dieks observes, our experience, and our experience of the passage of time in particular, gives no support to a metaphysical theory of time involving the succession of a definite set of global

---

[259] James 1890.
[260] Reichenbach 1924.

simultaneity planes. A different choice of such hyperplanes would make no difference to local experiences. As Dieks puts it we could "scrap the term 'simultaneity' from our theoretical vocabulary" and "no problem would arise for doing justice to our observations".

But it seems to suggest also that a metaphysical theory of time involving the succession of global nows is unnecessary for becoming to be real. In the previous sections we considered the possibility that cosmological considerations would perhaps secure absolute time picking on a preferred frame of reference even though the symmetry of inertial observers holds locally. Now if our experience of the passage of time is objective becoming experienced, and this experience is fully local, then whether cosmological considerations establish an absolute global time or not seems irrelevant to it. Then we studied quantum mechanics to see if it is compatible with thesis of the symmetry of local observers. But now it seems that it was irrelevant too. The symmetry of local observers is a thesis about something that is not local, namely the simultaneity of distant events, or the global now. It says of the latter that it is frame-relative and so unfit to ground ontological differences such as the difference between what has already come to existence and what hasn't. Very well so, if our experience of the difference between those two ontological states, and the transition between them, is local and so invariant under different choices of the frame of reference in which we are accounting for it.

Of course, to say that our experience of the flow of time and becoming is local, and to say that the flow of time and becoming is local, are not the same.

There is still the question whether we can account for becoming exclusively in terms of the invariant structure of Minkowski spacetime, without the succession of global nows. Such a theory would necessarily be a combination of presentism with an extensionless—pointlike—view of the present. It won't strike us as a surprise if that theory would involve a partial temporal ordering of events drawing on the invariant lightcone structure, relative to the apex of the past and future lightcones—present confined to a single point. We have to see if drawing on such a partial ordering of events we can make sense of the passage of time bringing about an ontological change, namely, coming to be.

In other words, we have to prove Gödel wrong in the first sentence of the passage we quoted from him. We have to show that it

is not true that "The existence of an objective lapse of time…means (or, at least, is equivalent to the fact) that reality consists of an infinity of layers of "now" which come into existence successively".

Stein's critique of Putnam's argument and the question of the relativizabiliy of existence

Historically, the theory that the present is local and becoming is to be accounted for in terms of the lightcone structure has been first advanced in the context of an objection to Putnam's formulation of the argument[261] from the conjunction of the relativity of simultaneity and the principle of no privileged observers to the unreality of becoming, namely, that Putnam's argument rested on a confusion. Howard Stein, one of the first and most influential critics of Putnam's argument, claimed that Putnam committed an elementary mistake in his argument: he was complaining about the lack of a relativistically meaningful answer to a question which itself wasn't relativistically meaningful. It is instructive to start the presentation of the local presentist case with the discussion of this objection to Putnam.

Putnam formulated the conflict he saw between the special theory of relativity and becoming by arguing that, if STR is true, there is no observer-independent answer to the question "Which events have become real?".

To be more precise, instead of the reality of events, Putnam talked of propositions about them having a truth-value. For our present concern, these two formulations of the matter can be treated as equivalent. In the article Putnam assumes that propositions have a truth-value only if the events to which they refer are real.

Putnam concluded that STR contradicts presentism because the question concerning which propositions have or have not a truth-value can be answered only relative to one's frame of reference, and that violates the principle of no privileged observers. He wrote:

> Why should a statement's having or not having a truth-value depend upon the relation of the events referred to in the statement to just one special human being, *me*?[262]

[261] Putnam 1967.
[262] p. 246.

Stein, however, protested that demanding an answer to the question which propositions have a truth-value in the relativistic context is an instance of the confusion of demanding a relativistically meaningful answer to a question which itself cannot be stated in a relativistically meaningful language:

> …"having or not having a truth value", in this question, must be understood classically to mean "at a given time"… but "at a given time" is not a relativistically invariant notion, and the question of definiteness of truth value, to make sense *at all* for Einstein-Minkowski space-time, has to be interpreted as meaning "definiteness at a given space-time point (or event)" – to be vivid, "definiteness for me now". The "Privileged Observer" (or, rather, privileged event) is – in effect – named in the question, and therefore has every right to be considered germane to the answer. Putnam's objection has an exact analogue, whose inappropriateness is plain, in the pre-relativistic case; namely, the question "why should a statement's having or not having a truth value depend upon the relation of the events referred to in the statement of just one special time, *now*?"[263]

Stein, an advocate of objective becoming, doesn't seem to worry much about the lack of an advancing global now, a conclusion of Gödel's and Putnam's argument that uncontroversially follows from their premise, the truth of the special theory of relativity. He writes:

> [I]n Einstein-Minkowski space-time an event's present is constituted by itself alone. In this theory, therefore, the present tense can never be applied correctly to "foreign" objects. This is at bottom a consequence (and a fairly obvious one) of our adopting a relativistically invariant language – since, as we know, there is no relativistically invariant notion of simultaneity.[264]

Stein is so relaxed about the relativity of simultaneity because he believes that for the purposes of objective becoming, a local present

---

[263] Stein 1968, p. 15. Stress in the original.
[264] Ibid, p. 15.

will do perfectly well. He proposes a partial temporal ordering of events relative to a point of reference, to be vivid, the event here-and-now, in terms of the lightcone structure, and regards the event here-and-now as the locus of the transition from futurity and ontological indefiniteness to pastness and ontological fixity.

But how this idea of reducing the present fairly dramatically to a single spacetime point combines with the idea that the present is the locus of coming to be? The latter is an ontological concept. Can an ontological concept have a meaning relative to a spacetime point? An ontological concept must refer to something objective.

Kurt Gödel, for one, thought that it is not even a matter for disputes that ontological concepts are not relativizable. He wrote in 1949 that "The concept of existence…cannot be relativized without destroying its meaning completely."[265] It seems obvious enough that if there is a difference between what exists for you, and what exist for me, it is not meaningful to say that either of us is right and the other is wrong, and it is not to be seen why either of us should have more say on the matter, than the other, then the difference between subjective appearance and objective reality comes to nothing.

Simon Saunders reacted to Stein's criticism of Putnam in a similar vein. He argued that the two situations which Stein described as "exact analogues" are different in a very important respect. It the pre-relativistic case definiteness was understood relative to an *intersubjective point of reference*, the *now*, whereas in the relativistic case the point of reference is not intersubjective any longer. Definiteness relative to the intersubjective now is replaced by "*definiteness for me now*", and, as Saunders puts it, "this changed situation is…simply no longer hospitable to presentism" anyway[266]. He also argues that if the question is whether an alleged ontological difference between events can indeed be real, and if no observer is allowed to be privileged over others, then intersubjectivity is a minimal condition for objectivity. If the cause of Stein's dissatisfaction with Putnam's argument is that Putnam has failed to formulate the problem in a relativistically meaningful language, then this can easily be repaired, since

> the requirement of intersubjectivity is certainly relativistically meaningful.[267]

---

[265] Gödel 1949, p. 558.
[266] Saunders 2002, p. 8.
[267] Ibid., p. 11.

I will argue that this last point of Saunders's about the requirement of intersubjectivity, namely, that it is relativistically meaningful, is true only with an important qualification. His previous point about intersubjectivity, namely, that it needs to be required as a precondition for objectivity, needs to be qualified too, accordingly. His first point about the requirement of intersubjectivity, namely, that in the "changed situation" to which Stein refers, i.e. when "already" is understood relative to a single event, rather than to a simultaneity plane, as a point of reference, is "simply no longer hospitable to presentism", because it is no longer hospitable to any kind of intersubjectivity about the present, is false.

Stein, when criticizing Putnam, wrote that "the question of definiteness of truth value, to make sense at all for Einstein-Minkowski space-time, has to be interpreted as meaning 'definiteness at a given space-time point (or event)'—to be vivid, *'definiteness for me now'. The 'Privileged Observer' (or, rather, privileged event) is – in effect – named in the question*, and therefore has every right to be considered germane to the answer."[268] This formulation of the point suggests, and I think misleadingly, that one can go back and forth between *relativity to an observer*, on the one hand, and *relativity to a spacetime point*, on the other.

If that was so, Saunders would be right that Stein's construal of temporal determinations fails to meet the requirement of intersubjectivity. But I would like to argue that these two kinds of relativity should be distinguished. Once these two kinds of relativity are distinguished we will see that relativity to a single spacetime point is not completely destructive to the Gödelian-Saundersian requirement of intersubjectivity. I will argue also that as long as the requirement of intersubjectivity is restricted to the set of observers in respect of which *it is reasonable* to require the intersubjectivity of becoming or existence, that is, in respect of which the requirement of intersubjectivity is relativistically meaningful, relativity to a spacetime point is compatible with that reasonable requirement of intersubjectivity. We may secure intersubjectivity with respect to all the *relevant* observers. So maybe relativizing the concept of existence to a spacetime point can be done without destroying its meaning completely.

---

[268] My emphasis.

The intersubjectivity of A-determinations and becoming with respect to the relevant observers

Stein defines A-determinations relative to a localized present, in terms of the lightcone structure. Now the lightcone structure of spacetime is intersubjective. The invariance of the speed of light secures it that all observers that share the same local present, i.e. all observers whose worldlines intersect at a point, which is, for them being at that point, *here-and-now*, agree about which spacetime points are lightlike separated from them, irrespective of the direction and speed of their motion, and also about the basic causal properties of the events that are inside and outside of the two lightcones.

On the special theory of relativity the speed of light is a limit for the propagation of all conceivable causal influences, therefore every event which might have had a causal influence on the event that is happening here and now, are inside and on the surface of one of the lower lightcone. Similarly, all events that the event here and now might possibly influence causally are inside or on the surface of the upper lightcone. Events that are outside the two lightcones (spacelike separated events) cannot have any causal connection with the event here and now.

Now the temporal determinations of the A series, being past, being future and being present, have to have an ontological significance, they correspond to ontological closedness or definiteness, to openness or indefiniteness, and to the limiting case between the two, which is the locus of the transition form one ontological state to the other, respectively. Can the temporal determinations defined in terms of the lightcone structure have this ontological significance? It seems that they can.

Ontological openness and closedness are of course not causal properties. Nevertheless, there is an intuitive interrelation between causal properties and the ontological significance of the temporal determinations of the A series. All events on which what takes place in the present may exert a causal influence are future. Now supposing that the causal order of the world allows for the event that takes place in the present to be underdetermined by the past, yet for some events in the future be fully determined by the present event (together with some spacelike separated events), then those events must be ontologically open. All events that might have had a causal bearing on what takes place in the present are past and must be ontologically

fixed to have an unambiguous causal influence. It was so in the pre-relativistic case, and the same interrelation between causal properties and temporal-ontological determinations is preserved in Minkowski spacetime if we identify the past with the lower, the future with the upper lightcone. The intersubjectivity of the causal structure of spacetime represented by the lightcone structure secures it that the proposed definition of A-determinations relative to a point of reference will be in line with the causal order from every observer's point of view.

'Having already become real' is then defined as a two-place predicate, or a relation between two spacetime points. y has become real, relative to x, if, and only if, y is in the lower lightcone of x. Similarly, not having become real yet, is a relational property, too: y has not become real yet, relative to x, if, and only if, y is in the upper lightcone of x. This definition of becoming makes no reference to entities or properties that depend on the choice of a frame of reference, so it is relativistically invariant.[269]

Indeed, the observers whose worldlines intersect in the point that has been chosen to be the point of reference relative to which temporal determinations are defined agree about what is past, what is present, and what is future, on the above definition. But what about the myriad of other observers?

When we inquire about the judgement of other observer's about what we, whose worldlines intersect here and now, judge to be the past, the present, and the future, first we should make it clear at which point of their worldlines we want the opinion of the others. It seems obvious enough that we do not want their consent to our ascriptions of temporal determinations to events at just any point of their worldlines, as it should be obvious in the Newtonian case, as I have argued against McTaggart, that the fact that someone tomorrow

---

[269] Stein 1991. This definition fits our specific concern, i.e. defending libertarian freedom. As far as freedom is concerned, our concern now is securing an ontologically open future that contains events which are up to us. It is the upper lightcone of the supposedly free agents where those events are to be found. If no effect can travel faster than light then no agent has any power over any events outside his upper lightcone anyway, whether or not those events are ontologically open. Given that ontological openness or closedness is a matter independent of the causal organization of events, and that it is causal relations that pick out the events from the upper light-cone over which we want to secure the agent's power, and any further hypothesis on what events are causally accessible for the agent would breach the generality of the account, we should require that the whole of the upper lightcone be part of the ontologically open future, and we should not require it to contain any spacetime point outside the upper lightcone.

would say of my dinner tonight that it was in the past, although it is future for me now, doesn't mean that incompatible determinations are predicated of the same event, or that an essential intersubjectivity between observers is lacking, so A-determinations cannot correspond to anything real. Even in the classical case, intersubjectivity concerning judgements about temporal determinations is required only in respect of some specific moments of observers.

The obvious, and notoriously pre-relativistic, answer to the question at which point of their worldlines we want the opinion of the others would be that we want the opinions they would give *now* (at the point of their worldlines in which they are now). But that answer is meaningless in Minkowski spacetime. There is no such thing as the *now* extended in space that could cross-sect their worldlines.

Consequently, there is no point in their worldlines which could be identified as the one at which it is relevant to ask their opinion about our ascriptions of temporal determinations. Then it is not relevant to ask them at all. Intersubjectivity is *not* a relativistically meaningful requirement without qualification. It is meaningful only in respect of observers whose worldlines intersect in the point of reference. But then, given that all inertial observers whose worldlines cross each other here and now agree about what is past, what is present, and what is future, on the above definition, we have *the intersubjectivity of all relevant observers*.

So the situation with Saunders's three points is this.

It is admitted that the requirement of intersubjectivity is relativistically meaningful, but only with this qualification.

It is not reasonable to require, concerning their judgements about whether an event has already become real or not, the intersubjectivity of observers that are spacelike separated from us (from the point of reference), because simultaneity is not relativistically meaningful (temporal modifiers are not transferable spatially). The justification of our uninterestedness in the opinion of observers timelike separated from us (from the point of reference) is the same as in the classical case. So we have a limited pool of relevant observers, those whose worldlines intersect here and now (at the point of reference). The intersubjectivity, as a precondition of objectivity, of judgements about existence should be required, but it should be required only with respect to the relevant pool.

The requirement of intersubjectivity is fulfilled within this relevant pool of observers.

Admittedly, it is a striking feature of the ascription of temporal determinations that has been proposed that there are spacetime points to which no temporal determination is ascribed. The localized version of presentism entails that there are events, those that are spacelike separated from the event that is taking place here and now, in respect of which there is no truth about whether they have already become real, or not, that is, there is no truth about their existence. It is not just that we don't know whether spacelike separated events are real, or not. It is not an epistemic matter. It is that, objectively, they do not have either the property of existence, or that of nonexistence. Is it acceptable? Doesn't it, as Gödel said, destroy the meaning of existence completely?

I think not. I think the best way to get used to this idea is to keep in mind that it is not the case that the existence or nonexistence of some events is objectively indefinite *simpliciter*, rather, it is the case that their existence or nonexistence is indefinite *relative to a point of reference*. As Stein emphasized it, this is not a completely new situation. In Newtonian spacetime the existence or nonexistence of events is also relative to a point of reference, but in that case, the point of reference is the global now, and relative to that, every event either exists or not. In Minkowski spacetime, however, such an extended point of reference could only be chosen arbitrarily. So the point of reference is reduced literally to a single point. A point, a much humbler object than a global simultaneity-plane, *cannot be relevant*, as a point of reference, to the existence of every event. It is irrelevant, as a point of reference, to the existence of those events that are spacelike separated from it. It doesn't mean that the ontological status of those events would not be definite, relative to spacetime points that are relevant, as points of reference, to their existence. The irrelevance of the event here and now, as a point of reference, to the existence of events that are spacelike separated from it, is intuitive. If we are here and now, the events outside both our lower and upper lightcones are completely inaccessible for us. They cannot affect us by any means, and we cannot affect them by any means. As if they weren't there. We know that they are there because they are points of worldlines of objects of the past of which we are informed, and on the future of which we may have an influence. To make it vivid: When we talk, I see and hear you. What I see and hear is your past. When I say

something to you, I send out signals to a future self of yours. I know that there is a gap between the two (even if it is extremely small if you are near), and I assume that the gap is filled in by you, that is, I assume that your worldline is continuous between the point from where the last signal reaches me here and now, and the point which the fastest of the signals which I send out from here now will reach first. It would be totally pointless, though, to ask which is the point until which the part of your worldline that fills in the gap has become real by now that I see and hear your past, and send visual and audio signals to your future. The new idea that local presentism requires us to get used to is that this question is not just without any practical relevance, but that this question is without meaning. I don't find it difficult to swallow at all.

The Augustinian worry concerning the ontological exclusivity of the present revisited: Doesn't local presentism actually lead to solipsism?

Augustine was deeply worried by the question that, if the present time was to be attributed with ontological exclusivity, wasn't it too thin to contain everything that existed. What troubled him was present's lack of temporal extension, without which, it seemed to him, it was not fit as the container of reality. Now we are advised by Stein and Dieks that we should let go of present's spatial extension, as well. What we are left with is a single spacetime point. On presentism, the present is the locus of coming to be, the locus of existence in a sense in which the past and the future is not. Can that locus of ontological exclusivity be pointlike? Isn't local presentism trivially absurd?

To this problem my attention was drawn by Simon Saunders when he supervised me for a year. He hinted that local presentism was tantamount to solipsism. This hint is present in his printed work, too. In his 'Tense and Indeterminateness' he writes:

> I am constructing presentism as an ontological doctrine; that all that is is what is now. I take it that "what is now" is a 3-dimensional space, along with sundry events; moments in the careers of objects. It is therefore a time-slice of space-time. If we are to take the position seriously, as a philosophical thesis, it is a public space – for nothing else exists, this is the whole of reality; and I take it that

> solipsism is not a serious position in philosophy. So we had better all agree on how this public space is to be defined.[270]

The target of this passage is local presentism. The locus of existence must be "a public place", otherwise we fall captive of solipsism. That is why there is no question about that presentism should be understood as the theory of a unique advancing global hypersurface of coming to existence, and local presentism is not even considered. And, therefore, presentism falls, because, as shown by Gödel and Putnam, we will not "all agree on how this public space is to be defined".

I argued in the previous section, defending Stein's construal of A-determinations against Saunders's charge that choosing a single spacetime point as a point of reference is not giving up totally on intersubjectivity, that the point-present is in fact a public space, public to all observers whose opinion is relevant to deciding which events have already come to existence relative to this point of reference; all observers whose worldlines intersect at this point.

But as I sit in my chair here and now, I know perfectly well that it is impossible for any other observer to cross my worldline here and now. I am simply too solid for that. My claim about the intersubjectivity of judgements about the ontological status of events relative to the event here and now can be true only of ghostlike observers who can penetrate each other. Or, alternatively, it can be understood counterfactually: if this point of spacetime was occupied by another observer, moving differently from the way I do, he would make the same judgements as I do. Or, yet again, it could be understood approximatively: no other observer will actually cross my worldline but if one comes really close, our ontological judgements will be very similar. Maybe it is enough for that claim of mine about intersubjectivity to be meaningful. As a matter of fact, however, I am alone in the spacetime point I am occupying. So if the present is just this point, then I am the only inhabitant of it. Now, if, on Stein and Dieks's advice, I commit myself to local presentism, do I commit myself to the theory that I am the whole world, nothing exists but me?

Not at all.

---

[270] Saunders 2000, Section 2.

The reason why local presentism does not lead to solipsism is that the local presentist is simply not committed to the view that only the local present exists. This is because the temporal ordering of events, according to his theory, relative to the local present, is only partial.

There is indeed ontological exclusivity involved in local presentism, but it is applied only to the events which are temporally ordered. It is true that as far as all the events go that are temporally ordered relative to the event here and now, my point of reference, only the event here and now is real. Events in the lower lightcone have already passed by, events in the upper lightcone are yet to happen. But the theory remains silent of the events spacelike separated.

The local presentist is committed to the view that the spacetime point he occupies is a locus of coming to be, but he is not committed to the view that there is only one locus of coming to be. He is only committed to the view that the infinitely many loci of coming to be do not combine together to a spatially extended locus, like a spacelike hypersurface, of coming to be. Or, which is the same, he is committed to the view that it is meaningless to ask which extensionless loci of coming to be combine together to form an extended one.

We have already discussed how a local presentist would account for his communication with someone else, another observer, with whom he cannot share "a public space" of existence, for the simple reason that they cannot overlap. Perhaps it is useful to revisit this situation to appreciate the fact that the local presentist is not a solipsist at all.

Suppose you and I meet some place, right now. Our worldlines are going parallely. We engage in communication. Suppose we exchange smiles that travel with the speed of light. Do I have any reason, being a local presentist, to think that from the two of us only I am real, and you are not, or not in quite the same sense as I am?

I see your smile, and I know it is from a moment of you which is on the edge of my past light-cone, so, according to my theory, it is no longer real. I smile back instantly, knowing that it will be received by a moment of you which, according to my theory, has not become real yet, since it belongs to my future lightcone. My causal interactions are with moments of you which are not real in the same sense as I am now. Does it mean that I am interacting with someone unreal, or less real than myself?

Not as long as I stick to the hypothesis that the moment of you that smiled on me and the moment of you on which I smiled back are connected in a continuous way by intermediary moments of you, preserving your personal identity. For then the moment which is not real *yet* and the part that is not real *any longer* must be connected by something that bridges the realms of *not-yet* and *not-any-longer*. Along the worldline that connects them there must be you coexistent with me.

Of course, there are other conceivable explanations for my experience of you. Your images do not belong to the same temporally extended entity, the temporally extended you is just my construction, I am dreaming, I am a victim of some deception, I am hooked up in the Matrix, etc. But this is the simplest and the most plausible explanation, and nothing in my local presentism contradicts it.

Neither does the special theory of relativity. It only says that it is not meaningful to ask which point of the—from me spacelike separated—part of your worldline that connects the two moments of you, the emitter of your smile and the receiver of mine, is you, coexistent with me.

Local presentism, instead of a monolithic four-dimensional picture of the world, in which all events are on a par with respect to their reality, canvasses a world with a flowing time, a time that flows locally. The local flow of time is an A series, presentism applied locally. On this picture, time flows locally everywhere. It is just that the local currents cannot be combined together to form a spatially extended river.

Summary and conclusions of the chapter

In this chapter I wanted to defend libertarian freedom against arguments to the effect, essentially, that libertarian freedom cannot be real, because it requires the future to be open in a sense in which the argument shows it cannot be open.

In the beginning I wanted to set aside two arguments to this conclusion, the logical fatalist argument as discussed, for example, in Aristotle's *De interpretatione*, and McTaggart's argument for the unreality of time, so that I can concentrate on the one which worried me the most, the argument from the special theory of relativity to the unreality of an ontologically open future (or the unreality of becoming).

For the special theory of relativity to be a real threat to the libertarian conception of freedom,

(1) the special theory of relativity must be true,

(2) from the truth of the special theory of relativity it must follow that becoming is unreal, and

(3) it must be true that the libertarian conception of freedom is tenable only if becoming is real.

Therefore, a defence strategy can have three main defence lines, consisting of arguments that can be advanced against (1-3), respectively.

I think all three lines may work. Yet, as I argued earlier in chapter two in relation to Kant's idea of exercising libertarian freedom in his timeless noumenal world, the third one would place a significant burden on the libertarian theorist's shoulder. The metaphysics of libertarian free persons in a monolithical blockworld would be a really tough work to develop and it is far from obvious how could one account for rational decision-making on the ground of what has been experienced in the past within such a frame. Since I believe in becoming I did not pursue this line of defence, but I would if I was forced by reason to accept that the special theory of relativity is true and that it implies the unreality of becoming.

The discussion of number one of the possible defence lines started with the recording of the fact that the special theory of relativity has an empirically equivalent alternative, Lorentz's theory. I argued that the real difference between the two is at the level of metaphysics. Lorentz and Einstein agree that the measuring rods and the clocks of inertial frames in relative motion are deformed relative to each other. Lorentz insists that only the spatial and temporal co-ordinates measured by undeformed rods and clocks represent real space and real time, even though it cannot be known which frame of reference has undeformed measuring devices, whereas Einstein insists that all reference frames capture space and time, even though it is plain that the co-ordinates they are using for this purpose are (mostly) deformed. This minimal difference between the two theories has grave consequences. Lorentz's theory is hospitable to presentism understood as the theory of an objective advancing global three-dimensional hyperplane of simultaneity, the global now, which is ontologically exclusive, in reference to which becoming, and ontological openness/fixity can be made sense of, while Einstein's isn't. It should be understood that the choice between these two

theories are not made on empirical grounds, it is rather made on other principles of theory selection, such as simplicity and compatibility with other theories one endorses.

From that point on there are three main sub-lines within this defence line.

One is that perhaps only Lorentz's theory is compatible with quantum mechanics, Einstein's isn't. This is the line taken, for example by Karl Popper and John Lucas. They appeal to the correlations between spacelike separated quantum events such as those involved in EPR-like scenarios, and claim that they can be accounted for only if absolute simultaneity is assumed. As it is presented in the Appendix, the fate of this argument varies from one interpretation of quantum mechanics to the other, considering the interpretations favoured by considerable portions of the physics and the philosophy of physics community. So it is best considered to be an open question.

The second is that even if the local symmetry of inertial observers is unaffected by quantum mechanics, maybe global considerations concerning our cosmological view of the whole universe prescribe a preferred frame of reference, deciding the choice between Lorentz and Einstein in Lorentz's favour. One of the champions of this view is Quentin Smith. Several arguments have been considered for this claim, the weightiest among them perhaps the ones drawing on the Friedmann-Robertson-Walker model of the spacetime of our universe and on one of the empirical evidences in support of Big Bang cosmology in line with the FRW model, the cosmic background radiation. Several important counterarguments were however considered too, mostly by Kurt Gödel, Simon Saunders and Dennis Dieks. I think the question whether a preferred frame of reference can be identified on cosmological grounds should be considered to be an open one, too.

The third is simply that even in the lack of support from either quantum mechanics or cosmology the preference for Einstein's theory rather than Lorentz's can be questioned. It can be argued that if compatibility with other physical theories does not select one rather than the other, compatibility with metaphysical theories we value should be taken into account among the principles of theory choice. If what has been argued in the fifth chapter is right, namely, that rationality is based on the capacity of choosing freely in the libertarian sense, then, if one theory can be squared with libertarian freedom

while the other is not, then the former is legitimately chosen on that account, since we value highly the theory that our theories are the results of rational reflection.

On the whole, line number one, I believe, is a promising line of defence, if one is forced to rely on it. I, personally, would regret if I would have to rely on this one, because I am attracted to the simplicity and intellectual elegance of relativity theory, although I am aware that this is not a very solid principle of theory choice.

For me, it would only be a fall-back strategy, since I hope defence line number two, that is, arguing against the thesis that the special theory of relativity would be inhospitable to objective becoming, should be successful in itself.

This favourite defence strategy of mine was initiated by Howard Stein who proposed that for the present to make sense in relativity theory it should be confined to a single space-time point, given that the notion of simultaneity is not relativistically meaningful, and taking advantage of the fact that future and past relative to a point of reference can be defined in terms of the lightcone structure, which is relativistically invariant. This proposal was lent support by Dennis Dieks's observation that our phenomenal experience of the passage of time and becoming, which is perhaps our most solid motivation to maintain the A theory of time and the belief in becoming, is fully local, insensitive to what goes on in spacelike separated points, therefore, insensitive to the relativity of simultaneity. My contribution to the argumentation for this view was only that I tried to defend it against Simon Saunders's largely Gödelian counterargument, drawing on the observation that intersubjectivity in respect of matters of existence is a relativistically meaningful requirement, and claiming that this requirement local presentism cannot meet. I hope to have shown that the requirement of intersubjectivity in matters of existence is met by local presentism, as long as the requirement is restricted to those observers in respect of which it is relativistically meaningful, given the relativity of simultaneity. The resulting picture of time is a myriad of objective local currents, bringing about the ontological change an advocate of becoming should want to be real, which however do not combine to form an objective extended flow.

We can conclude that the argument from the special theory of relativity is not likely to bring down libertarian freedom.

With this I conclude the second larger unit of the thesis which was concerned with the question whether—for our best scientific knowledge—we can have genuine (objectively possible) alternatives to choose from when we choose freely how to act. The answer definitely is that nothing that has so far been discovered rules this out.

Two more questions are yet to answer, both of which concern the libertarian conception of control, i.e. whether it can be conceived coherently, and whether it can be rational. This is the topic of the next chapter.

8 Can We Have Control, Especially Rational Control, Non-Causally?

The problem of freedom: distinguishing some determined actions from reflex movements, or some cases of underdetermined activity from random involuntary movements, in a morally relevant way

In the fourth chapter I argued—challenging Daniel Dennett's arguments to the contrary—that the worry that blaming deterministic wrongdoers is unfair is motivated by a U-condition for responsibility, which our moral intuitions endorse. The U-condition required that an action for which the agent is to be held morally responsible should not be an event necessitated by a set of jointly sufficient causal conditions for which the agent was not responsible.

Whether or not some of the causal conditions that figured in that jointly sufficient set were internal to the agent, or even to his deliberative faculty, made no difference. If the action was necessitated by a choice, and the choice was necessitated by a psychological state, including beliefs and desires or whatever is invoked in deliberations, of which the agent was not responsible, then he was not responsible for the action.

The intuitive wisdom behind this judgement seems to be that even though the contribution of his deliberative faculty might have been a necessary condition for the production of an action, we do not take the agent to be genuinely active with respect to this action if it was causally necessitated by conditions in respect of which he was clearly passive. If he was passive in respect of these causal conditions, then he was moved helplessly along a course of action-production by forces about which he could do nothing. Although the action produced is undeniably his, in one sense of the word, since it was performed by him and it was a consequence of a choice or decision issued by the deliberative faculty of his mind, the relationship that binds him to his action, that makes the action his, is not that of an author or originator and his product. On a closer look, what he did proves to be a mere sequence of happenings, and he is related to these happenings as the mere scene in which all these took place.

Thus the agent is passive in respect of his action. He was passive with respect to the states and events that were jointly causally

sufficient for the action to occur, and causal necessitation transferred passivity from the necessitating cause to the necessitated effect.

If this transfer principle is correct, then in a deterministic world we are passive in respect of every action. In a deterministic world for any occurrence there can be found a set of jointly sufficient causal conditions, in the early universe, say, in respect of which we are clearly passive.[271] If causal necessitation transfers passivity, no occurrence really deserves the name "action", since being passive in respect of one's action sounds like a contradiction in terms.

If there is a hope to avoid the conclusion that determinism entails passivity in respect of everything that occurs, it is by denying the principle that passivity is transferred through causal necessitation. (I will call it Transfer Principle 1, or TP1 hereafter.)

But can this principle be reasonably denied? I cannot think of any set of premises with stronger or more immediate appeal than the principle itself, which would entail it. The principle of the transfer of passivity through causal necessitation strikes me as something that spells out directly what we mean "passivity" and "causal necessitation". The idea of causal necessitation is that some events are just not floating loosely in the flow of all events. They are tied together by productive powers so that the occurrence of one is the product of the occurrence of the other (or a set of others). If activity (or passivity) is a relation between an agent and an event that is characteristic of the agent's role in the event's coming about, i.e., whether it is *originative* or not, then the relation that an agent bears to a necessitating cause he bears also to the effect, given the way that the occurrences of the two are tied together. Nothing is necessary for the coming about of the effect over and above the necessitating cause. So there is no room for the agent's activity or origination to step on the scene if it wasn't already present in the cause.

I see no way to deny TP1, and therefore, I see no way to avoid the conclusion that an *action* that truly deserves this name, i.e., in respect

---

[271] Here I am assuming that it is uncontroversial that we are passive in respect of the conditions that obtained in the early universe, which is very obvious as long as we assume that the phenomenal flow of time is real, entities come to exist at some time and seize to exist at some later time, and that it applies to persons, too, and consequently we must be passive in respect of the physical state of the early universe, since it obtained at a time when we haven't yet existed. In line with the results of the previous chapter, I assume that becoming is real.

of which the agent is active, should be causally underdetermined by states and events in respect of which the agent is passive.[272]

But what about underdetermined events? Can one be active in respect of them? The worry is that if someone's φ-ing was causally underdetermined then that he φ-ed instead of something else, instead of an event of a different kind that could also have happened all causally relevant factors being equal, was something that turned out that way just at random. Is it any more reasonable to hold one active in respect of a random occurrence than of something that was causally necessitated by factors beyond one's reach?

Perhaps the ordinary strong conception of agency—which I take to be the idea that we are capable of doing things which are not merely happening to us, either in the sense that we are made to do them by powers beyond our control, or in the sense that our doing them is just a random brute fact popping out of nothing—is incoherent. Either an occurrence is causally determined, or it is not, the field of options is exhausted by these two. If both exclude that the occurrence have the property that an agent is active in respect of it in the ordinary strong sense, then there is no such thing as agency in the ordinary strong sense.

This is essentially what Hume suggested, adding that if one is considering a more relaxed way of making sense of agency, not wanting the impossible, then determined events are the fitter candidates for being actions in this more relaxed sense. If the causal chain that leads up to the event goes through the agent's deliberative faculty, so that the occurrence is necessitated by a choice issued in that faculty (which is nevertheless necessitated by factors in respect of which the agent is passive), then, at least, the outcome reveals

---

272 Klein discusses the principle on pp. 98-100 of *Determinism, Blameworthiness and Deprivation* (1990) and says she finds it appealing *in the abstract* yet she does not find it persuasive when *applied to concrete cases* like the production of action by reasons which just occur to the agent. She uses "trying" as an example of something that the agent does which might arise deterministically from desires and beliefs in respect of which the agent is passive, which nevertheless is hard to think of as a case of passivity. I take it as a recording of a conflict between intuitions. As far as "trying" is concerned, our intuitions differ. I don't find it difficult to think of tryings as cases of passivity if they are necessitated by beliefs and desires in respect of which we are passive. But I don't want to insist on the transfer of passivity through causal necessitation in just any sense of the word "passivity". What is really important for my arguments to come later in this chapter is the transfer of passivity in the sense relevant to moral non-responsibility. Perhaps our intuitions are more in line as far as this weaker transfer principle (to be introduced two pages below) is concerned.

something of the agent, and the same cannot be said of a random occurrence. If we sort out an appropriately chosen[273] set of jointly sufficient causal conditions (in respect of which the agent is passive) into two subsets, one external, one internal to the agent, then we may even say that the outcome is something that the agent, being the kind of agent he is, reliably produces as a response to certain stimuli, so it is informative of the agent's character, as Hume emphasizes it in the eighth section of the *Enquiry*. So if we want control over what is happening to us, even if this control cannot be more than just a passive function of our character, we should want our actions determined rather than random.

So maybe in one sense passivity is not transferred through some cases of causal necessitation, after all. These cases are when the causal chain goes through the agent's character (in a specific way).

We are also advised by the Humean that holding someone responsible for actions that have arisen from his character deterministically is fair and normal, much more than holding someone responsible for whatever happens to him at random.

This I would contest, however. Even if I defer from defending TP1, I would insist that the sense in which a deterministic agent can be said to be active in respect of any occurrence (some weaker sense than the ordinary strong sense of agency), is not a sense which would ground moral responsibility. So I would insist on a weaker version of my original transfer principle (TP1' hereafter), which is that *moral non-responsibility* is transferred through causal necessitation. I think I have strong arguments in support of TP1'.

Fischer and Ravizza in *Responsibility and Control* (1998) attack TP1' head on. They think the principle can be easily sinked with counterexamples which are essentially cases of overdetermination.

This is one of them: Some meteorological events make it the case that an avalanche destroys a military camp at the foot of a snowy mountain. Betty was not responsible for the meteorological events, the meteorological events causally necessitated the death of the soldiers in the camp, so if TP1' were good, Betty would not be responsible for their death. But Betty is an officer of the army

---

[273] Of course, we may identify the set of jointly sufficient causal conditions somewhat arbitrarily, that is why I refer to an "appropriately chosen" set. If the world is deterministic, then, for any occurrence at time t, there is a set of jointly sufficient causal conditions at any time t', t'<t. An "appropriately chosen" set is one that obtains at a time t' such that at t' the agent already exists and has a character.

defending a pass in the mountain, and precisely in the moment when the meteorological factors would have caused the avalanche anyway she blows up a glacier to trigger the same avalanche to wipe off the camp of the enemy army preparing to take over the pass. Her blowing up the glacier would have successfully triggered the avalanche alone, and would have killed the enemy soldiers (also) in the absence of the meteorological causes. So Betty is responsible. So TP1' is false.

Maybe TP1' doesn't hold as it stands. But cases of overdetermination are absolutely irrelevant to our topic and to the relation between a cause, which is a mental state, and an action, which is an effect of that mental state, that TP1' was meant to express. No doubt, an agent can be responsible for an action causally necessitated by his state of mind even though he was not responsible for being in that state of mind, provided that he has other means apart from that state of mind to make it the case that the action happens, and the action is overdetermined by these means and by the state of mind in question. Can compatibilism be vindicated by appealing to this possibility? No way. The interesting question is whether non-responsibility transfers through the necessary causal relation that a determinist posits between a mental state of the agent before the action and the action, supposing that that causal relation is all that there is to be said about how and why the action took place. Of course, if the agent's irreducible self brings it about that the action occurs, say, via "agent causation", although his reasons would have causally necessitated the action anyway, then he is responsible for it, even if he wasn't responsible for having his reasons. But this is not a way for the compatibilist to make his case. If the compatibilist is to argue that the agent is responsible for the action although he wasn't for the mental state, he will do it by appealing to some specific features of the causal connection between the two and not by imagining that the agent might have controlled the occurring of the actions by some overdetermining other means. Fischer and Ravizza themselves do it this way. To TP1' we could add a clause to fend off cases of overdetermination, but, having regard to the obvious irrelevance of such cases, I omit it.

As far as the Humean point about the moral relevance of random occurrences is concerned, it is clear that we do not hold anyone responsible for random involuntary movements of his limbs, like a tremor, for example. But, I think, it is equally clear that we are not holding anyone responsible for reflex-movements either, if he had

nothing to do about the occurrence of the event or state that triggered it. The responses produced by the agent to environmental stimuli necessitated by his character are very much like reflex movements from a moral point of view.

Or are they? The responses to environmental triggers now considered are produced in a very complex, though mechanical, way, and one specific feature of this response-production is that the outcome carries the stamp of the agent's character. Is it then appropriate to assimilate the outcomes of such action-producing mechanisms to simple reflex-movements?[274]

As far as moral responsibility is concerned, I think it is. I think it is clear that we should not want to blame or punish anyone for his character. Blame or punishment is due for what one does, or fails to

---

[274] In my view compatibilism is essentially the business of trying to specify conditions, which, if they are met by an internal action producing mechanism, make it the case that passivity is not transferred through the mechanism after all, even though the mechanism is like Kant's turnspit: given the input, in respect of which the agent is passive, the outcome, the agent's action is fixed. Fischer and Ravizza, for example, suggest that if the action producing mechanism is reason responsive (meaning essentially that, had external circumstances provided the agent with some different reasons to consider, the action produced by the mechanism could have been different), the necessary and sufficient condition for the agent's being responsible for the action produced by the mechanism is his previous acceptance of the mechanism as "his own" (meaning essentially that he, at least tacitly, agreed to consider himself responsible for whatever is produced by this mechanism), and whether the mechanism works in a causally determined way or not is irrelevant. This suggestion strikes me as profoundly counterintuitive. If determinism holds, the agent is a victim of causal factors in respect of which he is passive in his whole life including the episode or period when he started to accept as adequate the retributive behaviour of his environment toward the deeds he produced in a certain way (canvassing alternatives that seemed possible from the internal perspective, considering reasons, and choosing). Why would the fact that he was caused to accept responsibility for whatever he is caused (in a certain way) to do make it the case that he is objectively responsible for it? Fischer and Ravizza's answer seems to be essentially that assuming this we can account for how an agent is responsible in Frankfurt-type scenarios in which we clearly hold him responsible although he is (supposed to be) robbed of alternative courses of action. If this is the philosophical rationale of accepting this *prima facie* counterintuitive proposal, then it shouldn't motivate us to accept it, because, as we have seen in the fourth chapter, Frankfurt-type agents do have an alternative in the required sense, so their being responsible is not difficult to account for without the Fischer-Ravizza proposal. Their other condition on an action-producing mechanism's being responsibility entailing, i.e. reason-responsiveness, is also inadequate, if my arguments in chapter 5 are sound. Mechanical reason-responsiveness is a mark of non-rationality if the conclusion of chapter 5 is accepted. Rationality cannot be mechanical. Requiring that and agent's action could have been different had his reasons been different is insufficient for rationality. What should be required is that his action could have been different even with his reasons held fixed.

do, not for what one is like. To a bad character it may be normal to react with dislike, but no one with a bad character deserves moral blame as long as the bad character is not manifested in wrongdoing. But if it is true that the wrongdoing is mechanically produced by the character in the presence of some environmental trigger about which the wrongdoer has nothing to do, then the distinction between being blameworthy or being morally clean is just a matter of pure luck. The counterintuitive conclusion that pure luck decides whether one is blameworthy or not can be avoided only by denying that one can be responsible for actions that arise deterministically from one's character of which one is not responsible, and this is tantamount to lumping deterministically produced actions together with reflex-movements in the relevant respect.

The only way to resist this conclusion is to claim that we are responsible for our character. I am sympathetic to this idea (defended enthusiastically by Sartre for example), but certainly not if we are assuming determinism and if by "character" we mean a set of internal-to-the-agent causal conditions of action production, in respect of which the agent is passive. For an agent to be responsible for some part of his character it would be required that it would be up to him now, or that it would have been up to him some time in the past, to have or not to have that part. But that would require that his having that part of his character meet the U-condition. But that requires libertarian freedom. (Sartre was a libertarian.) So assuming responsibility for character is not an option for the compatibilist.

Or isn't it? There are philosophers who argue that we do not have to have any choice about our character to be morally blameable for it.

Robert Merrihew Adams offered the following example.[275]

> Suppose you have just realized that you are ungrateful to someone who has done a lot for you—perhaps at great cost to herself. Far from responding to her sacrifices with love and gratitude, you have made light of them in your own mind; and if the truth be told, you actually resent them, because you hate to be dependent on others or indebted to them. Surely this attitude is blameworthy. Must

[275] Adams 1985. My attention was drawn to this article by Martha Klein who generously read and commented an early version of this chapter in 2004, for which I am very thankful. Adams's article is also discussed by Fischer and Ravizza (1998, pp. 255-9).

> we assume that you have caused it, or let it arise, in yourself by actions that you have voluntarily performed or omitted?

Then Adams looks for possible voluntary actions or omissions in the past that might have led to the ungrateful state of mind. He considers the proposal that the ungrateful person could have fighted against his ingratitude but he omitted it. But is it reasonable, he asks, to assume that the omission must have been a voluntary one to lead to a blameworthy state of mind?

> You have not begun sooner to struggle against this ingratitude. But it would not be correct to say that you have thereby voluntarily consented to the bad attitude. For voluntary consent, as ordinarily understood, implies knowledge; and you did not realize that you had a problem in this area. How then can you be blamed for not having fought against your ingratitude?…You should have known of your ingratitude. Why didn't you? Presumably because you did not want to recognize any shameful truths about yourself—because at some level you cared more about having a good opinion of yourself than about knowing the truth about yourself. And that's a sin too, though not a voluntary one. Thus the search for voluntary actions and omissions by which you may have caused your ingratitude keeps leading to other involuntary sins that lie behind your past voluntary behavior.

This example, however, leaves me entirely unconvinced. Don't we experience many times that little children can be ungrateful, too? Possibly they have not yet realized that it is better to be grateful than ungrateful, or possibly they do not yet have an even remotely adequate conception of what gratitude means. Is their ingratitude sin, too?

Adams's story is complicated by the phenomenon of failing to recognize and fight ingratitude in consequence of a general attitude that consists of preferring a good opinion of oneself to the truth about oneself. He says that is a sin, too. Now, can't a child have that attitude? I seem to remember a time when I saw my father as mythical hero. Only the fathers of my best friends were remotely comparable to him, but of course only remotely. Why was it so? Was

it so partly because I subliminally thought that having a very good father meant being valuable? I preferred seeing him supernatural to seeing him as he was, a strong, brave, smart and original man—one of many. I also seem to remember a time when I assumed that I couldn't be happy without being good at most things I tried. On the look back, I think I somehow subliminally suspected that I would not be very good at football, so I avoided the playgrounds where the other boys played football, so that I wouldn't have to face it. I preferred a good opinion of myself to the truth. Was that a sin?

I think it wasn't, and neither is childish ingratitude before the child first becomes conscious of it.

Now how should we account for the fact that we are willing to blame the grown-up for his ingratitude and not the child? I think the most plausible explanation for the difference between in our attitudes in the two cases is that we assume that an adult, fully equipped with powers of reflection and having enough experience, has already encountered occasions when his ingratitude (or his preference of good self-esteem to the truth) must have occurred to him and he had the opportunity to reflect on it, distantiate himself from it, and finally do something about it. So, at bottom, we assume it was up to him to amend his attitudes or leave them as they were before. So we assume that that they were left as they were before was a voluntary omission.

In his example Adams is referring to a prior attitude (preference of having a good opinion of oneself to knowing the truth of oneself) that caused it that the person in question has never encountered an opportunity to reflect on the attitude in respect of which we are now asked to consult our intuitions whether we hold him blameworthy for it (the ingratitude). The prior attitude is invoked in order to canvas a scenario in which failing to change the latter attitude was an involuntary omission. Of this prior attitude he says that it was already a sin. Now is it *because* the prior attitude was already a sin that the attitude being concerned (whose remaining unrevised was necessitated by the prior attitude) is a sin too? Or would the latter be a sin even if the former wasn't? Adams's remark that caring more of a good self-esteem than of the truth is a sin seems to suggest that he thinks that its being a sin is explanatory of why we should consider the case of ingratitude he is discussing a sin. Surely, Adams thinks that the history of how one acquired a bad attitude or failed to change it is relevant to the question whether one is blameable for it. If it was planted in him, or if his becoming conscious of it was blocked, say, by

a superscientific manipulator, then we would certainly exculpate him. If it is important that the prior attitude was a sin, then why shouldn't we enquire about the person's history that led to his having that prior attitude, to make sure that that is really a sin? If history was relevant in one case, why would it not be in the other? Are we going to refer to a third attitude, which is also a sin, to make it plausible that the second attitude is a sin, although the person had nothing to do about it, given the third attitude? If so, where will we stop?

There is a long tradition in the Christian religion, now present predominantly in Protestantism, holding that our *nature* is sinful, independently of our having to do about it. This tradition goes back, through Luther and Calvin to the older Augustine, and is grounded scripturally in Paulian theology (especially in the epistles to the Romans and the Ephesians). But the tradition of non-reformed Christianity, especially Eastern Christianity, has always been suspicious of this idea. It is no accident that Pelagius, the fifth century Christian champion of the libertarian conception of virtue and sin, who was condemned as heretic on Augustine's incentive for contradicting the doctrine of original sin and related teachings of the Church, was later largely rehabilitated by the Orthodox churches of the East. Protestant theologians often accused the Catholic Church of having converted to Pelagianism after the Council of Trent in the sixteenth century. Intuitions on cases like the one presented by Adams have always diverged and this divergence of intuition seems to be largely responsible for a great doctrinal divide within Christianity. Adams's intuitions in this question are on the Protestant side, mine are on the Orthodox side. I accept that it may make sense to talk of involuntary character traits as "sinful" of "fallen" but not in the sense that conveys culpability.

There was a regress that was looming in Adams's referring to a prior condition that he declared a sin, namely, preferring a good esteem of oneself to the truth, which in Adams's example was responsible for the person's inability to reflect on his ingratitude.

I suspect Adams suggested that that prior condition was a sin, because he shares my intuition that the person wasn't *more responsible* for not being grateful as he was responsible for his longstanding inability to reflect on his ingratitude, and he was not more responsible for his inability to reflect on it than he was responsible for having the prior attitude that made him unreflective of it.

I think Adams shares this intuition because I think it rests on a simple principle that should be appealing even to most compatibilists. I call it TP2. TP2 is this: *If an event is an inevitable consequence of some environmental conditions given the agent's character, then he is not more responsible for this event than he is responsible for having his character.* I think if a compatibilist is motivated to claim that we are responsible for our character, it is because he wants to rescue moral responsibility even though he accepts TP2. I think most compatibilist are much more likely to accept TP2 than my earlier transfer principle TP1'.

But TP2 is an ambush for compatibilists. If TP2 is granted by the compatibilist, then there is no way for him to escape the conclusion that a deterministic agent is not responsible for his character, because anyone who accepts TP2 has to accept TP1' as well.

Once TP1' is accepted, non-responsibility for the character (or for anything) evidently follows from determinism, unless we assume responsibility for the initial conditions set out for the causal evolution of the universe in the Big Bang.

The compatibilist can only challenge TP1' by insisting that there are some necessitating causal chains that do not transfer moral non-responsibility. I don't see any other candidates for this special status than the causal chains that go through one's character. So the compatibilist would essentially need to defend the thesis that if a necessitating causal chain goes through one's character (some specific way) then moral non-responsibility is not transferred through it.

So the only way for an agent to be responsible for having (the relevant part of) his character, although his having it was necessitated by causes for which he was not responsible, would be if the causal chain leading from those causes to his having it went through his character. Through the character he already had at the time the causal process took place.

But if TP2 is accepted then he cannot be more responsible for this newer edition of his character than he was responsible for the older edition that he already had when the new edition was causally produced.

Now we may ask how he could be responsible for this older edition of his character if it was a necessary consequence of causal conditions for which he was not responsible (like the physical state of the early universe). The answer must be the same: He could be responsible for it if the causal chain leading up to it from those causes

went though his character. The character he already had at the time this older edition of his character was produced.

And so on.

The compatibilist has embarked on a recursion that will stop when it reaches an early edition of the agent's character for which he is clearly non-responsible, e.g. the character he had right after his birth. At that point this recursive account for responsibility for one's deterministically had character collapses. (Supposing that we accept at least one bit of Pelagianism. One of the doctrines Pelagius famously denied was that children who die unbaptized are excluded from salvation.)

So a compatibilist accepting TP2, who nevertheless thinks he can resist TP1' because he thinks that some causal chains that go through our characters do not transfer moral non-responsibility, is bound by TP2 to accept that we cannot be responsible for a deterministically formed character.

Now if we combine this conclusion with TP2 again, then the result is that we cannot be responsible for anything that is a necessary response to an environmental stimulus given our character. So we are not responsible for anything that is necessitated by causes of which we are not responsible, whether or not the causal chain leading up to it goes through our character. So the compatibilist has to accept TP1', after all, if he accepted TP2.

But then he has to accept that we are not responsible for anything if determinism holds. To apply the language we used earlier in the chapter, our actions are just reflexes.

The compatibilist-libertarian debate can be presented as a rivalry between two philosophical projects that seem *prima facie* equally hopeless: saving something of agency by distinguishing some determined actions from reflex movements, and attempting the same by distinguishing some underdetermined actions from random involuntary movements, in a morally relevant way.

The compatibilist attempt to distinguish some deterministically produced actions from reflex movements, from the morally relevant perspective, necessarily fails, unless we deny TP2, the principle that no one is more responsible for an event that is an inevitable response to some environmental stimuli, given one's character, than one is responsible for having the character. But I don't see how TP2 could be reasonably denied. So this project seems entirely hopeless.

In this chapter we examine if it is really equally hopeless to try and distinguish some causally underdetermined actions from mere random involuntary movements. Can some of them be distinguished from mere randomness in a morally relevant way?

Distinguishing underdetermined activity from randomness in a morally irrelevant way: causal indeterminism

In the previous section we have come to the conclusion that causal necessitation transfers moral non-responsibility from the cause to the effect, so if the world is deterministic, our moral non-responsibility for the conditions that obtained in the early universe is transferred to our actions whether or not the causal chain leading to them goes through our personality (some way or another).

So either the world contains objective indeterminacy, or there is no moral responsibility.

But what if we suppose that the world is not deterministic? Then perhaps we have a chance to find events for which we can be held responsible among those that are not the deterministic causal products of conditions for which we are clearly not responsible.

Supposing that we are not creating ourselves responsibly out of nothing, there is a time early in the life of each of us, when what we are and how we are is none of our making. At that time we are not responsible for what we are and how we are.

So if there are events for which we are responsible, those events should not be the deterministic causal products of what we were and how we were at that time. Unless it is possible to conceive coherently of events that are not the deterministic causal products of what we are and how we are at the time, yet we are responsible for them, there is no responsibility.

Why? Is liberty from causal necessitation by what we are and how we are really a necessary condition for responsibility also at later times (later than the early phase of our personal development when we are clearly not responsible for what we are and how we are)? No. If we suppose that at later times we can be responsible for what we are and how we are, then it isn't. For it is not responsibility-diminishing if our action is a deterministic causal product of what we are and how we are at the time, if we are responsible for the latter.

Nevertheless, our responsibility for what we are and how we are now must have arisen somehow. And it couldn't have arisen from any

other thing than the capacity to responsibly bring about events or states of affairs that are not causally determined by what we are and how we are at the time when we bring them about, because deterministic causal consequences of what we were and how we were at the early time when we were not responsible for it would not yield responsibility, because non-responsibility is transferred through causal necessitation.

But this conclusion is a worrying one. For if an event is not necessitated by what we are and how we are at the time when we bring it about, then how do we bring it about, and how are we in control of its coming about in a way that can ground responsibility?

The causal indeterminist strategy is an attempted answer to this question. The answer is essentially this: Let what we are and how we are at the time cause the event we are to be held responsible for, but require that what we are and how we are causes it indeterministically.

For the sake of simplicity, instead of "what we are and how we are" I will use "reasons" as a shorthand. And I will use "choice" as a shorthand for "the event for which we are to be held responsible".

With these shorthands the causal indeterminist strategy is that reasons should cause, but should not necessitate, the choice. Control should come from causation, and freedom from the transfer of non-responsibility should come from the non-necessitating character of causation.

But what is non-necessitating causation? The idea, if not comes from, is inspired by quantum mechanics. In standard (von Neumann) quantum mechanics an event (such as an outcome of a measurement) is not necessitated by previous events and the laws of nature. Yet it is not true that it would be uncaused. Previous events do bear a causal relation to it. They do not have a determining influence on its coming about, but they do have a determining influence on the probability of its coming about. If quantum mechanics is the ultimate account of how physical events hang together causally, then, at least on its standard interpretation, there is nothing more to be said about the causation of physical events. Nothing is left out of the picture. Causation is just influencing probabilities.

On the standard interpretation of quantum mechanics in a measurement process the state of the quantum mechanical system indeterministically collapses into a state that corresponds to one of the possible outcomes of the measurement (process 1). The indeterminacy is objective. There is no physical fact that could be

identified as the necessitating factor for which the state of the system would collapse into one eigenstate rather than any other. Yet, it is not the case that the state of the system at the beginning of the measurement process, right before the collapse, would have no bearing on which of the possible results will obtain. The state of the quantum mechanical system before the collapse can be mathematically rendered as a linear combination of the eigenstates (the possible post-measurement states). This is a feature of the algebraic structure with which quantum mechanical states are represented. It is a vectorspace in which eigenstates form a basis, just like three orthogonal vectors form a basis in the vectorspace with which ordinary three-dimensional space is represented. So the pre-collapse state of the system is mathematically a mixture of the possible post-collapse states, in which the latter appear with different weights. Their weights correspond to the probabilities of each possible outcome to actually occur. The bigger is the weight of an eigenstate in the mixture, the bigger is the probability that the system collapses into that state in the measurement process. The pre-collapse state of the system is often referred to as a "superposition state" in which all the possible outcomes are present with different weights.

Now there are indeterminist theorists of free will who suggested that the choice being undetermined yet caused by what and how we are immediately prior to the choice may be analogous with the post-collapse eigenstate being undetermined yet caused by the pre-collapse superposition state.

The two most influential authors that have proposed this solution are Robert Nozick and Robert Kane.[276]

They suggested that whenever we make a choice between different courses of action supported by different and conflicting sets of reasons, before the choice we are in a state that is the "superposition" of, say, willing to do A for reason R_A *and* willing to do B for reason R_B, and when the choice is actually made the superposition state collapses into one of the clean states of which the superposition state was composed, willing to do A for R_A, *or* willing to do B for R_B. Nozick emphasizes the structural analogy between choices being indeterministically caused by complex mental states in which we are inclined to follow different incompatible courses of action and quantum collapses, leaving open the possibility that this is only an

[276] Nozick 1981, Kane 1996, 1989.

analogy and not identity, while Kane seems more willing to expressly identify the event of choosing with a quantum collapse taking place in the brain, and speculates that chaotic deterministic processes in the brain, which are very sensitive to minor disturbances is their initial conditions, may amplify the indeterministic microscopic quantum events into macroscopic consequences.

It is a very important feature of the account of both authors that they say there is nothing more to be said about how the choice was brought about than the description of the complex mental state in which the agent was prior to the choice, that indeterministically caused the choice.

Such an account can provide for two important things that are necessary for freedom in the responsibility entailing sense: a multiplicity of alternative objective possible futures, and control over which one of them will occur. The former is secured by the assumption that the pre-choice mental state (the superposition state) is objectively compatible with a range of possible choices, the latter is secured by the choice's being nevertheless caused probabilistically by the pre-choice mental state.

Yet I don't think this is what it takes for a choice to be free in the responsibility entailing sense. If there is nothing more to be said about how the choice was brought about than the story about how it was indeterministically caused by a mixture of conflicting reasons, then I don't see why such an indeterministic causation should transfer moral non-responsibility less than if the relation between the cause and the effect was deterministic.

Peter van Inwagen wrote about the pre-choice superposition state, which allowed for both A and B to occur, contained both R_A and R_B, and indeterministically caused A, rather than B, that it

> did not *have* to cause [A]. Moreover, since it did not have to cause [A], and since it alone caused [A], [A] did not have to occur. But then did the agent have any choice about whether [A] would occur? ... Once [it] has occurred, then *everything* relevant to the question whether [A] is going to happen has occurred. After that we can only wait and see.

> In a perfectly good sense, it is going to be a matter of *chance* whether [A] occurs....[277]

Given that on the causal indeterminist account everything relevant to the question whether A (rather than B) would happen has already occurred when the pre-choice superposition state occurred, there is no fact to point at, which would make the agent *more responsible* for doing A (rather than B) than he was responsible for being in the pre-choice superposition state. Doing A (rather than B) is distinguished from merely random events, since, arguably, an event's being indeterministically caused is not the same as being random, but this distinction is *not relevant morally*. If there is nothing more to be said about how it happened that the agent did A (rather than B) than whatever is contained in the description of the pre-choice superposition state, then, supposing that the agent was not responsible for being in that state, his moral non-responsibility in respect of being in that state is inherited through indeterministic causation by the action this state indeterministically causes.

Robert Kane in *The Significance of Free Will* assumes an objection to his theory that invokes a situation similar to the one sketched by van Inwagen, except that a multiplicity of identically prepared agents are involved in it. Timothy O'Connor discusses this objection and Kane's answer to it in *Persons and Causes*:[278]

Consider an imaginary set of agents being in the same state of mind prior to a choice, in which some predictable proportion of them chooses one way and the rest the other. Suppose you are one of them, and suppose that one of the two alternatives is really hurtful for someone. Now, on the assumption that everything relevant to the coming about of the choice is contained in the pre-choice state, isn't it just a matter of luck whether you happen to be in the subset that does the hurtful thing, or not? Our intuition is that mere chance cannot decide who is blameable and who is not.

Kane's responded to this imaginary scenario by claiming that this scenario is misconceived, because exact sameness is not a predicate that could be applied to mental states. He writes:

---

[277] Van Inwagen 1983, p. 144. Van Inwagen uses different notations appropriate to the example (an example we will discuss a little later) he is discussing in the passage. I adapted them to our present context. The stresses are those of the original text.
[278] Kane 1996, p. 172; O'Connor 2000, pp. 40-41.

> This is how free will is related to the uniqueness of persons. ... Each life history is unique and cannot be exactly the same as any other if the psychological histories involve indeterministic processes, as they must do for free will. ... With indeterminate efforts exact sameness is not defined. Nor is exact difference.[279]

But this response will not do. O'Connor discussing this move by Kane protests, rightly, that Kane departs from the way he thus far followed in interpreting quantum mechanics when he tries to avoid this objection to his account by claiming that predicates such as exact sameness cannot be applied to pre-choice mental states. His previous identification of the pre-choice mental state with a quantum mechanical pre-collapse superposition state was dependent on a realistic interpretation of the latter, involving that it is a perfectly well defined, and in this sense determinate, physical state, even though it is indeterminate what results will obtain if some of the observable properties of systems being in this state are measured. If it was impossible to conceive of the sameness of such states, then all statistical predictions of quantum mechanics would be meaningless, and quantum mechanics would not get off the ground as an empirical science.

I would add to this that even if Kane was right in claiming that there is no such thing as two persons being in exactly the same pre-choice mental state, this is not "how free will is related to the uniqueness of persons". The moral of the imagined scenario is not dependent on the conceivability of there being an actual multitude of identically prepared agents, or even on the conceivability of identically prepared but numerically different agents. A counterfactual multiplicity of situations where the same unique person with the same unique psychological history chooses differently would do just as well.

If you happen to actually realize a situation in which you don't do the hurtful thing, although your doing it was equally possible given your whole pre-choice history, and if your pre-choice history contains everything that there is to be said about why you choose the way you choose, then you are just lucky not to realize a situation in which you do the thing, and by no means better, morally speaking, than you would be otherwise.

---

[279] Kane 1996, p. 172.

So this attempt of Kane's to fend off this objection fails, whether free choice is thought to instantiate a form of physical process conforming to a particular interpretation of quantum mechanics, or if the latter is used merely as a structural analogue.

If mere chance is not to decide who is responsible and who is not, you are not any more blameable, supposing that you are one of those who do the hurtful thing, than those who don't. So if you are to be held responsible for this action indeterministically caused by your pre-choice mental state, then your moral responsibility for it *must* consist in your responsibility for your pre-choice state of mind. But this conclusion means that causal indeterminism fails, because causal indeterminism can account for moral responsibility only if there are cases when we are responsible for a choice emerging from a state for which we are *not* responsible. If there aren't such cases, what this example seems to show, then it is hard to see in virtue of what a grown-up would be more responsible for anything than a newborn is, who is clearly not responsible for what and how he is right after his birth.

At many places Kane seems to suggest that some specific details of the agent's mental history leading up to his choice may secure that his choosing to do A rather than B was not a matter of chance in the morally relevant sense, after all, even if it is the case that both were possible given his pre-choice mental history that indeterministically caused his choosing to do A, and that nothing over and above his pre-choice mental history was relevant to the question whether he would do A or B.

I think it is impossible to find the details that could do this work for him.

If the agent is responsible for how he chooses, then it is so either in virtue of a property of the choice itself, that secures responsibility irrespective of his pre-choice mental history, or something about the pre-choice mental history is also necessary for his being responsible.

On the causal indeterminist account there is no likely candidate for the property of the choice itself that would secure responsibility for it, irrespective of the mental history. According to the causal indeterminist hypothesis, the choice was indeterministically caused by the agent's pre-choice mental history, and nothing else had any bearing on it. We have seen that on such assumptions we are not more responsible for an effect than were already responsible for its indeterministic cause. So however we identify the crucial components

of the agent's mental history that we hope to generate responsibility—may them be the efforts of his will[280] to sort out his conflicting aspirations, or a relatively stable comprehensive network of neural connections linking different areas of the brain determining how an image induces a thought, a thought a memory, a memory a desire, a desire an aspiration, and the like, and how these all provoke, amplify, modify or inhibit each other, and how a synchronicity emerges from all this, possibly capable of controlling behaviour[281], or previous choices that had an effect on how this network is hooked up[282]—the critical property these components need to possess in order that the agent be responsible for whatever they cause (indeterministically) is the same: the agent must be responsible for them. So we are embarked on a recursion enforced on us by the transfer of non-responsibility through indeterministic causation (with the assumption that nothing else has a bearing on the effect apart from the indeterministic cause), which is inevitable to reach past stages of mental history for which the agent is clearly not responsible. And at that point our hope to establish responsibility by invoking the specific role these components of mental history play in the indeterministic causal production of the choice brakes down.

I can think of only one more possible property of some of these components of our mental history that may be suggested to secure that we are responsible for what is indeterministically caused by them. This possible property is that they constitute *us*.

Maybe the comprehensive neural network of the last paragraph (the "self-network") is just what we are. Maybe there is no gap between it and us that should be bridged by our being responsible for it in order to secure that we are responsible for what it causes. O'Connor seems to take Kane to essentially suggest this solution when he refers to the self-network.[283]

Now even if we adopt this reductionist approach to ourselves (to which Kane himself perhaps does not subscribe) the original problem remains unsolved. If the link between us (the self-network) and our action is indeterministic then how is it not just a matter of chance, morally speaking, that our action turns out to be one of the possibilities left open by what we are, and not any other? It seems we

---

[280] An "effort of the will"—Kane 1996, p. 126.
[281] A "self-network"—p. 140.
[282] "Self-forming actions"—p. 37.
[283] O'Connor 2000, p. 40.

would be much better off with determinism. The relation between us and our action would be a determining one then, and then this question would not arise.

But then, of course, other questions would arise. How is it that a robot is not morally responsible for what it does, although what it does is a deterministic causal consequence of how it is programmed, and how it is programmed is essentially what it is? Surely, it is its programmer who is to be held accountable for what the robot does, not the robot. The question whether we are responsible for what we are is a meaningful one. It is conceivable that we are not responsible for it even though there is no gap between us and it, it is just us. And then we are not responsible for what it causes deterministically for the reasons stated earlier.

It seems that Kane wants libertarian free self-forming actions partly because he wants us to be responsible for what we are, at least to some extent. He clearly thinks that that is a necessary condition for our being responsible for what it causes deterministically. (That is why he thinks the U-condition, or Ultimate Responsibility condition, to use his terminology, is a necessary condition for moral responsibility.) But if our responsibility for what we are, secured by its being partly the result of self-forming actions, is a necessary condition for our being responsible for its deterministic causal consequences, then so it is for being responsible for its indeterministic consequences, as well. For, as we have seen, if the indeterministic causal story exhausts what there is to be said about why a choice occurred, then it transfers moral non-responsibility just as a deterministic story would. However, if those self-forming actions are themselves indeterministically caused, the way causal indeterminists suggest, then it is unclear how we are in control of them in a responsibility-entailing way if we, i.e. the previous edition of ourselves, bear only an indeterministic relation to it. And so it is unclear how we are responsible for what it brings about, i.e. the new edition of ourselves.

Robert Nozick seems to have an interesting proposal here. He suggests that we do bear a non-ambiguous relation to a self-forming choice that relieves the worry that, insofar as the indeterministic cause of our choice leaves open alternative possibilities, it is a matter of chance, in the morally relevant sense, that we choose one way rather than the other. He says that in such choices a "conception of oneself and one's appropriate life" is also chosen by the same token, and the

choice is thus not "a random brute fact", because "it will be explained as an instance of [that] very conception". Rather than being a chancy one, the choice is a "self-subsuming" one.[284]

So the trick is that the indeterminism involved in the causal production of those choices leaves room for genuine alternatives, yet the fact that we choose to act one way and not the other is not just a matter of chance, in the morally relevant sense, because it is uniquely connected to a conception of ourselves we adopt by that very choice.

This suggestion strikes me pretty much like one of the extraordinary adventures of Baron Münchausen in which he pulled himself out of the water by his hair so that he wouldn't drown. Of course he would have to be already out of the water to pull anyone out. Reference to a later state when he is out and thereby capable of applying a pull on someone in the water does not help, even though that later state would obtain immediately when the pulling is successfully performed.

Genuine alternatives are there because there is no unique connection between any particular possible outcome of the choice and the *pre-choice edition* of what we are. And exactly for the same reason, it is a matter of chance how we choose, in the morally relevant sense, if there is nothing more to be said about why we choose so, as it is assumed by causal indeterminists, Nozick included. The fact that there is a unique connection between how we actually chose and the *post-choice edition* of ourselves does not change this. It was a matter of chance, morally speaking, that we chose that edition of ourselves by choosing the way we did.

I conclude that causal indeterminism fails to provide for what it promised. The proposal that we should picture genuine free choices as being indeterministically caused by what we are and how we are at the time of the choice (assuming that nothing else is relevant to the question why a particular choice occurred) provides for real alternatives, and provides for some sense of control. In this sense choices indeterministically caused by reasons are distinguished from both determined and random events. But they are not distinguished from random events in a morally relevant way, because there is no answer to the question how it is not a matter of chance that an agent chooses to act one way rather than the other, if both were live alternatives given what he was and how he was prior to the choice.

---

[284] Nozick 1981, p. 300.

There is a trade-off here between the alternatives' being "really live" and the choice's not being a matter chance. If the probability of an alternative to occur is close to 1, given what and how the agent is prior to the choice, then, of course, in a perfectly good sense, it is not a matter of chance if that alternative obtains, even though the causal link between the pre-choice state of the agent and his choice is indeterministic. But then the other alternatives are not really "live". The extent to which the alternatives are really live is exactly the extent to which the choice is chancy.

I agree with Barry Loewer[285] in that if the work libertarian philosophers want quantum mechanics to do for them is that they want free choice to be identical with, or structurally analogous to, the collapse of the superposition state into one of the eigenstates, then quantum mechanics will not help them to make sense of freedom in a way that grounds moral responsibility in a way that compatibilist freedom doesn't.

As far as the transference of non-responsibility is concerned, indeterministic causation is different from deterministic causation in just one way: the former can perhaps allow for something more to be said about how the effect was brought about, while the latter doesn't. If there is more to be said, then the agent may be more responsible for an action or choice than he was responsible for what he was and how he was right before the choice occurred, *in virtue of this extra*. If there is no such extra, then moral non-responsibility is transferred just as if the link between the pre-choice state and the choice was deterministic.

Another attempt to distinguish free choices from random events: they are non-random, because they are explicable in the light of reason

The causal indeterminist way of trying to distinguish some undetermined choices from mere randomness, in a morally relevant sense, does not succeed. But perhaps there is another way.

David Wiggins thinks there is. He writes,

> This objection [if a choice is undetermined then it is random] is question-begging. One cannot prove that determinism is a precondition of free will by an argument

[285] Loewer 1998.

> with a premiss tantamount to 'everything is either causally determined or random'. This is simply too close to the conclusion, that whatever is undetermined is random. That is what had to be shown.[286]

Wiggins believes that the premise is false and so is the conclusion. But he sees a challenge here, which has to be answered by the libertarian. He continues,

> But in the form of a challenge, it may appear that at least something of the objection can stand. If an event is underdetermined, if nothing excluded an event of different specifications from taking the event's place, then what does it mean to *deny* that the event is random? What is it to be justified in ascribing the action identical with the event (or comprised by the event) to an agent whom one holds *responsible* for that action? In the unclaimed ground between the deterministically caused and random, what is there in fact to be found?[287]

This is the challenge that causal indeterminism could not meet. Wiggins does not give a very detailed account of how he thinks this challenge can be met. He gives some important hints though. He says that

> [the agent's] possible peculiarity as a natural thing among things in nature is only that his biography unfolds not only non-deterministically but also intelligibly; non-deterministically in that personality and character are never complete, and need not be the deterministic origin of action; intelligibly in that each new action or episode constitutes a comprehensible phase in the unfolding of the character, a further specification of what the man has by now *become*. … It may not matter if the world *approximates* to a world that satisfies the principles of neurophysical determinism, provided that this fails in the last resort to characterize the world completely, and provided that there are actions which, for all that they are causally

[286] Wiggins 1998, p. 290.
[287] pp. 290-1. Stress in the original.

> underdetermined, are answerable to practical reason, or at least *intelligible* in that dimension. Surely *these* are not random. They are the mark left on the world by conscious agents who have freedom.[288]

I take it that Wiggins suggests that what an agent is to do at a given time is, at least sometimes, when the action is free, causally free-floating, i.e. multiple alternatives are left open by the totality of causal factors. This creates a scope for practical reasoning. The outcome of the agent's being engaged in practical reasoning is that, from the causally possible courses of action, the one that actually takes place gets picked out. This process is non-causal by nature. There are no events or states, within or outside the agent's mind, to which the selected course of action would relate as an effect to its cause. *A fortiori*, the agent's reasons to act the way he does are not causes of his action.

It does not change much if some reasons are allowed to work as causes, and then some parts of practical reasoning can be translated into the language of causal explanation; as long as there are reasons which are not causes causal explanation is incomplete, and the gaps can be filled out with genuinely non-causal rational explanation. Assuming that this is what Wiggins wants, let us see if he can get it.

In the first chapter we shortly touched upon an argument offered by Galen Strawson to the effect that "being underdetermined but explicable in the light of reason" cannot be the property that ultimately delineates a third category of events, besides determined and random, that would be the realm of genuine freedom and self-determination. Strawson thinks that in order to have the kind of self-determination libertarians should want, which is different from setting up our mind randomly or determined by causes we have no means to control, we should be self-determined in respect of whatever it is that explains our choice. If it is reasons doing a non-causal work in the process through which our choice is formed, then we need to be self-determined to have those reasons. If self-determinedness means what Wiggins suggested it does, then the reasons that explain our choice we should have as a result of an earlier choice that was underdetermined but non-random because it

---

[288] pp. 293-4. Stress in the original.

was explicable in the light of reason. In the light of earlier reasons, of course, in respect of which we should be self-determined, too. Clearly it leads to an infinite regress.

So introducing a third category, "explicable in the light of reason", alongside determinedness and randomness, does not save us from ultimately falling victim of causes determining us without us having any say about that, or of blind chance. Saving us from both determinedness and randomness would require reasons to form an infinite hierarchical and temporal array (reasons higher in the hierarchy would have to have been adopted earlier), and that is not possible to have in our heads.

This regress argument resembles closely the "Rylean" regress argument we also discussed in chapter 1 to the effect that if freedom is understood as determinedness by the will, then freedom must be shallow or superficial in the sense that it can penetrate only a finite number of layers of the factors that causally determine our action. If we keep to the idea that control must be a kind of causal necessitation by already existing facts or events, then we cannot be "absolutist" about the requirement that what we do should not be determined by causes other than our will (causes with which we cannot identify, or with which we cannot be reasonably identified). Events determined by the will (by "the right kind of things"), do not constitute an interesting third category that would save freedom from the dilemma of determinedness by things other than the will (by "the wrong kind of things") and randomness. Of course, if we are not "absolutist" about freedom from determining factors in respect of which we are passive, then we can quit the regress at any point.

Strawson does not present his regress argument as a variation on Ryle's. But Strawson apparently thinks that the regress he points out is as bad, from an "absolutist" perspective, as Ryle's, and he thinks that the libertarian perspective is necessarily an absolutist one. I think this regress is indeed bad if we adopt *absolutism about rationality*, but not very bad, or not bad at all, even from the perspective of a quite ambitious libertarianism.

The most economic way to show why giving up on absolutism about rationality is not so bad for a Wigginsian libertarian is to present Strawson's regress argument against the background of Ryle's.

If Wiggins is right to claim that some events are undetermined yet non-random *because* they were rationally chosen, then we have three

categories of events: (A) causally determined events, (B) random events, and (C) events that came into being because they were chosen for a reason. Is it true that a theorist of freedom who is an "absolutist" about undeterminedness by causes other than the will (a libertarian) doesn't get any further with this typology than he does with the tripartite typology that was discussed by Ryle, in which we had (A) events that are causally determined by other things than a will (by "the wrong kind of causes"), (B) events that are random, and (C) events causally determined by a will (the "right kind of causes")? An infinite rational explanatory hierarchy of category C events in the Strawsonian case is just as much absurd as is an infinite causal sequence of category C events in the Rylean case.

The consideration that launched us on a regress in the Rylean case was that we thought we needed self-determinedness in respect of the will, in order that we could be truly and not just superficially self-determined in respect of what is determined by our will.

Do we really need self-determinedness in respect of our reasons, if rational explanation is how Wiggins suggests it is: non-causal, but exempting from randomness? Are we launched on a regress the same way?

The account of free action Ryle considered was this: An action is free if it was determined by a free will, and a will is free if it was determined by a (previous) free will. The comparable way of stating the Wigginsian account of free action would be this: An action is free, if it was determined by a free will, and a will is free if it was chosen for a reason, i.e. if it was a Category C event.

If we say only this, without requiring self-determinedness in respect of the reasons for which we choose to will something, it is already a richer sense of freedom than if we would simply say that freedom is determinatedness by the will. Suppose that the reason for which the will was chosen was random. It is different from the comparable Rylean case when the (second order) will itself is random, for the reason does not causally determine the will, so it is not the case that our action is a necessary causal consequence of a random occurrence. Suppose now that the reason for which the will was chosen was determined. Now, again, this case is different from the comparable Rylean case when the second order will is determined, because being chosen for a reason is *ex hypothesi* different from being determined by a reason, so freedom from determination is retained (we cannot fall victim of Walden Two-type cases, for example).

But are we really rational if reasons just pop up in our minds randomly? Are we truly autonomous if our actions can be explained with reference to reasons we are determined to have? The "unclaimed ground" between determination and randomness is very narrow unless we require what Strawson said we should require, that acquiring the reason for which we made up our mind should have also been a Category C event.

Yes, the unclaimed ground is narrow, but it is there. Doing something for a reason can be a self-forming action in Robert Kane's sense, if rational explanation is the way Wiggins suggests it is, regardless of the way we adopted the reason. It is not itself random, nor can it be traced back causally to any combination of random and determined causes. At least in this minimal sense, the requirements for libertarian freedom are met even if the reason for the action is determined, or if it is random. If we want deeper freedom, we can ascent to higher levels in the envisaged hierarchical structure of rationality without the risk that we cannot stop. We have the right to stop at any level. Level 1 was good enough for securing a non-empty sense of undetermined yet non-random self-determination. If the reason for which the action was done was adopted for a reason (that is level 2), that is even better.

Surely, an infinite hierarchy of reasons in which every single reason was adopted for another reason would be impossible. Someone might recommend that, instead of the infinite hierarchy, to vindicate absolutism about rationality, we should envisage a circular but large and coherent alliance of reasons, in which every single reason is explicable in the light of other reasons, like, according to the coherentist theory of epistemic justification, a big alliance of statements may justify each other, without any of them being justified in the foundationalist sense. But it doesn't help. Because the task is not only to make reasons intelligible in the light of others (timelessly), but also to account for how they came into being, how they were adopted. That is why we acknowledged earlier that the hierarchy of reason is also a temporal one.

But we shouldn't worry about that. Requiring that every reason should be adopted for a reason is not just impossible, but—as it was argued in the third chapter—it is also unnecessary.

Reasons are normally thought of as combinations of epistemic and normative elements. We desire something—that is the normative element, and we think that a course of action is a good means to

realize it—that is the epistemic element. With qualifications discussed in chapter 5, we could be happy to be endowed with the undetermined yet control-preserving freedom to choose only the normative elements of our reasons, provided that someone guarantees that the epistemic elements, what we think the relations are between means end ends, are always correct. So what freedom in respect of reasons essentially comes to is the freedom to choose desires. Yes, the libertarian maximalism about freedom clearly involves that we want freedom in respect of our desires. But not in respect of all of them, or not in respect *all of them at the same time*. We should not want to create ourselves out of nothing. Presumably there are good desires, encoded in our genes and in our culture that we can happily embrace. To give a banal example, I am sometimes told that I resemble my grandfather with whom I lived until I was five. They say it is not only looks, but patterns of behaviour, as well. Presumably, part of the explanation for this resemblance is that, through genetic or cultural inheritance, I have some of the desires he had. I was always happy to hear comments about our resemblance from our family, because I have a very good memory of my grandfather.[289] Strawson is right that it can never be the case that we have an infinite simultaneous array of desires in which each and every desire we have we have because we chose it for reasons that are also part of the array, but it doesn't mean that we cannot rationally revise any desire we feel necessary to revise. Strawson's argument doesn't guarantee immunity from rational revision to any desire we adopted for reasons that we no longer endorse, or acquired determinately, or randomly, for no reason at all. Were I frustrated by some patterns of my behaviour that make me like my grandfather, I could explore their motivational background, and try to revise the desires that motivate me to behave so, realizing that they conflict with other desires I value more. It would be hard work, I suppose, and it is possible that I would prove too weak to overwrite some of the old desires. But that weakness would have nothing do with Strawson's argument. So the scope of self-formation by way of rationally revising desires that is open to us despite of Strawson's regress argument can be really ambitious, even by libertarian standards.

[289] I was led to this point by Harvey Brown in a discussion in the early summer of 2004.

I believe Strawson considers the suggestion that libertarian free actions are chosen for reasons as a putative alternative to the causal determination–randomness dichotomy, which libertarians may try, but, he thinks, will fail to defend, for the reasons he is offering. The text, however, allows a reading according to which choosing for a reason is a special case of causal determination, when the cause is a reason. I presented Strawson's argument as if the former was the correct reading of it, as if it was targeted against the Wigginsian suggestion.

If I was wrong in so presenting the argument, if Strawson considers choosing for a reason as a case of causal determination, then, of course, he is right to claim that it will not lead the libertarian anywhere. If reason is suggested to be *the right sort of thing*, which the causal theorist of control should want to causally necessitate free actions, then the causal theorist is facing an infinite regress. But this is not a very interesting conclusion, since we have already learned it from Ryle. The Rylean argument was not sensitive to what the criterion for a cause to be of the "right sort" was. It was dependent only on the requirement that the right sort of thing, which is allowed to cause without diminishing freedom, should neither be random, nor causally determined by things of the wrong sort. Indeed, reason cannot be the right sort of thing in that sense. Nothing can. That was Ryle's point. Libertarians must construe control some other way than causal determination by the right sort of internal causes.

If Strawson's argument is meant to be an argument against the Wigginsian proposal, as I presented it, if we are not asked to assume that choosing for a reason is a case of causal determination, then it is not a successful argument. The Wigginsian proposal, if it otherwise works, does not run into a regress analogous to the Rylean one, for the reasons I have given.

I presented Strawson's regress argument as if he, for the sake of argument, granted that Wiggins can get what he wants, i.e. that rationally chosen events may be both undetermined and non-random, in view of being able to show that this doesn't lead the libertarian anywhere, because his absolutism about self-determination would launch him on a regress of reasons adopted for prior reasons. So granting to Wiggins what he wants leads to absurdity.

I objected against this argument on the ground that if non-random undeterminedness in virtue of rational explicability is granted to the

libertarian, then the libertarian is not necessarily motivated to ascend to higher levels of self-determinedness with respect to his reasons. Being passive in respect of (a good deal of) one's reasons is compatible even with an absolutism about the revisability of reasons.

It seems that once it is granted that rational explicability exempts from randomness, even if reasons do not cause what they explain, nothing absurd follows from that with reference to what, what has been granted for the sake of argument, could have been taken back. But maybe the opponent of the Wigginsian proposal doesn't need to argue against the proposal in such an indirect way. Maybe the impossibility of what Wiggins wants can be shown directly.

The work rational explanation is supposed to do by Wiggins is to secure that what we rationally choose to do is non-random, although reasons are not causes and do not causally necessitate what we do. That would be Wiggins' way to save libertarian freedom from both incoherence and irrationality by the same token. Can reasons possibly do this work?

To illustrate what that would mean, let me borrow an example from Peter van Inwagen.[290] Suppose the complete description of the relevant part of the mental state of a thief, upon facing the decision whether he should rob the poor box or not, consists of references to two conflicting belief-desire pairs. One (A) is the desire to keep the promise he made to his dying mother that he would lead a decent life, in combination with the belief that on the given occasion the course of action that would realize this desire is refraining from robbing the poor box. The other (B) is the desire for money, in combination with the belief that on the given occasion robbing the poor box is the adequate course of action that would realize this desire. Suppose he refrains from robbing the poor box. He would report (and suppose a criminal psychologist would confirm), that he has refrained from robbing the poor box *because* he had the first of the two desire-belief complexes (for Reason A). What can and cannot this *because* mean if it is to be a Wigginsian *because*?

Suppose a boson emitted millions of years before in the radioactive decay of an atom of a distant galaxy hit a serotonin molecule in the thief's brain just when the face of his dying mother occurred to his mind's eye right before he would have reached his hand for the poor box, and that this is why this belief-desire complex

[290] van Inwagen 1983, pp. 140-1.

issued in an action, and not the other one. Had that genuinely random quantum-mechanical process of emitting the boson in the depths of the past of a distant region of the universe happened a nanosecond later, the guy would have robbed the poor box. Of course, in this case both the thief and his psychologist would report that he refrained from robbing the poor box *because* he had Reason A to do so. In one sense, it is true. But, quite evidently, this *because* isn't worth more than the "because" of any *post facto* rationalization, since it means *only* that the thief had both the desire and the belief that make up Reason A, and that much is true. Surely it is not a Wigginsian *because*. We have to require more from rational explanation if the work we want it to do for us is to save action from randomness.

Now suppose that both Reason A and Reason B are realized in neurophisiologically describable brain states, and Reason B stimulates the neurons that are in the position to send the electrochemical impulse to the muscles in the thief's arm and shoulder that would realize his reaching out for the poor box, whereas Reason B inhibits them. Suppose that the inhibition is way stronger than the stimulus. In this case too, the thief would report (and his psychologist would confirm) that he refrained from robbing the poor box *because* he had Reason A to do so. Very probably, the introspective phenomenology of this *because* is not at all different from that of the *because* of the previous case. But, of course, objectively it is a very different *because*. It is the *because* that relates a determining cause to its effect. It is not the Wigginsian *because* either.

Are there other kinds of *because*?

We can suppose that reasons work in a non-causal way, even if we don't have an absolutely detailed account of how exactly. They are considered, weighed, they recommend courses of action, they incline us, but there is one thing they don't do. They do not cause. Or maybe they cause, they exert a causal influence on us, but they do not causally *necessitate*. Causal influence that falls short of necessitation is quite harmless. Of course, we want reasons to do more than just indeterministically cause the choice. We have seen that that is not enough for making the choice non-random in the required sense. We want them to do a genuinely non-causal work for us.

So what about the non-causal work reasons do in action production? Is that necessitation, or something less then necessitation? (Provided that it makes sense to talk of an event necessitating another without being its cause.)

If it is necessitation, then we are in trouble. Because then our catalogue of kinds of events again consists of three types: (A) events necessitated by the wrong kind of thing (a cause), (B) events that are not necessitated by anything, in any way, events that, to the extent their outcome is left open by influencing factors that, taken together, fall short of necessitating it, are random, and (C) events necessitated the non-causal way by reasons. The Wigginsian suggestion then is that a libertarian free action is a Category C event. Now, what if the occurrence in the thief's mind of the reason that necessitated the Category C event, our thief's action, is a Category A event? An example would be if it was the case that the state of the universe a nanosecond after the Big Bang, together with the laws of nature, necessitated that our thief has the non-causally necessitating Reason A at the time of his action. Then his action is a necessary consequence of the wrong kind of things, in this example the laws of nature and the physical state of the newborn universe. What if the occurrence of the reason that necessitated the Category C event, the thief's action, is a Category B event? An example would be if it was a consequence of a genuinely random radioactive decay that he has the non-causally necessitating Reason A, (or more realistically, if it was a consequence of a genuinely indeterministic quantum mechanical event that he remembered his mother at the time of his action, which caused that Reason A, which he had dispositionally, was triggered to occur to him). Then his action would be a necessary consequence of a random occurrence. He is not self-determining in refraining from robbing the poor box in either case. So Reason A is better to be a Category C event. So it must have been necessitated by a reason. Now ask the same questions about that reason, and we find ourselves embarked on a regress right away.

The upshot is that assuming that the Wigginsian *because* refers to *necessitation by reason*, which is non-causal, but nevertheless necessitation, the Wigginsian proposal to make sense of libertarian freedom fails because of the Ryle-Strawson regress argument.

Now suppose that the non-causal work done by reasons falls short of necessitation. Then our thief's action is random to the extent of the ambiguity left about it by all the factors that have an influence on it, causal, or non-causally rational, unless the thief, after all work those factors, including reasons, could do is done, somehow *himself* determines what to do.

The action is not decided by any*thing*. As far as *things* (events and states of affairs) that may have any bearing on the thief's action are concerned, his refraining from robbing the poor box is a free-floater in the flux of events. The only reason why it is not random is that the *person*, who then must be irreducible to the states of affairs and events that we might have though to constitute him and his flow of consciousness, determined it. The question of determination by the right or the wrong sort of thing, that led us to the Rylean regress, simply does not arise.

But now the worry is that even if we suppose that this determination by *persons*, contrasted to determination by *things*, can be conceived coherently, as far as it is the disambiguation of the ambiguity about the action left by every*thing*, reasons included, it seems patently non-rational, whether or not reasons work the causal way. As Strawson put it "the agent-self with its putative, freedom-creating power of partially reason-independent decision becomes some entirely nonrational (reasons-independent) flip-flop of the soul".[291]

The primitive power to create and its relation to rationality

Is this the end of the Wigginsian proposal? Yes and no.

Strawson's conclusion that the freedom-creating power of persons, if there is to be such a thing, must be an "entirely nonrational flip-flop of the soul" is wrong, as I will shortly argue.

But something like this is true. Strawson is right in claiming that Wiggins's proposal that some underdetermined events are non-random because they are explicable in the light of reason is wrong.

But something like what Wiggins proposed is true. Some events are explicable in the light of reason because they are underdetermined but non-random.

How would that be?

Strawson's argument was that, even without starting to enquire about what the "*because*" that binds a reason to an action it rationalizes might and might not be in Wiggins's proposal that there are genuinely free choices that are underdetermined yet non-random because they are explicable in the light of reason, this proposal can be shown to

[291] Strawson 1986, p. 54.

launch us on a regress similar to Ryle's, if we are absolutist about self-determination. We have seen that the libertarian is not bound to be an absolutist about self-determination in that sense. If it is assumed that rational choices can belong to a third category, neither determined nor random, then their presence in the process that leads to the choice may break the rule of the dichotomy of determinedness and randomness, and that is what a libertarian needs. A libertarian doesn't need to be committed to absolutism about rationality.

The problems started to arise when we started to enquire about what the Wigginsian *because* should mean. We couldn't find any sense to that *because* that would make it the case that it is its answerability to reason that makes a choice belong to a distinctive third category besides determinedness and randomness. It can be the case that rational choices belong to a third category, the realm of genuine activity, but an agent's being active in respect of his choice, in contrast to being a passive victim of either determinedness or randomness, cannot *consist in* his choice's being explicable in the light of reason. Supposing that more than one choice is consistent with his having the reasons he has (Case 1), it is unclear how it is non-random that he chooses one way rather than another that would also be answerable to, or intelligible in the light of, his reasons. Non-randomness with respect to this cannot consist in what was suggested by Wiggins. Supposing that only one choice is consistent with his reasons (Case 2), either he has the capacity to act irrationally or he hasn't. If he has (Case 2A), then the choice between acting rationally and acting irrationally cannot consist in what was suggested by Wiggins. If he hasn't (Case2B), then reasons bear a necessitating relation to the choice. Then this necessitating relation is either causal or non-causal. If it is causal (Case2Ba), then passivity (at least in the sense of moral non-responsibility) is transferred from the reasons to the choice, so, in order to be active in respect of his choice, the agent must have already been active in respect of his reasons. If it is non-causal (Case2Bb), then the situation is very much the same. There is no reason why passivity (moral non-responsibility) should transfer less through non-causal necessitation (if there is such a thing) than it transfers through causal necessitation. When we argued earlier for the transfer of non-responsibility through causal necessitation the only features of the relation between the cause and the effect the argument relied on were (a) that it was necessitating, and (b) it was completely immanent of the two related things, neither requiring nor allowing the

agent to add his contribution to the coming about of the latter. That it was necessitating in a special sort of way, i.e. causally, (if there are different ways of necessitating), played no role in the argument. If the agent doesn't have the capacity to resist what reason dictates, then the relation is between his reasons (his state of mind before the action) and his action is (a) necessitating, (b) immanent in the sense that nothing else is relevant from the agent's part to the coming about of his action over and above his having the reasons (being in that state of mind).[292] So he must have been active already in respect of his reasons, in order to be active in respect of his choice. Now, if being active consists in what was suggested by Wiggins, i.e. explicability in the light of reason, then we are facing the very same questions concerning the relation between the reason that explains the choice and the reason that explains the reason, and different answers to these questions will lead us to cases of types identical to those of Cases 1-2Bb. This is a regress with branches, and the regress plus the branches exhaust the field of options. All the branches and equally the regress end with the same conclusion: that activity, or non-random undeterminedness in respect of a choice, cannot consist in the choice's explicability in the light of reason. If that would have been Strawson's conclusion, then he would be right.

But he claimed more. He claimed that activity, if there is such a thing, must then be the intervention of a nonrational flip-flop of the soul. And that is not right.

We have discarded compatibilism because we found that events deterministically caused by reasons do not constitute activity in a

---

[292] If the truth to be told, I don't think this is a real option. I think if I were not committed to the view that reasons explanations are non-causal and non-necessitating, I would be committed to the view that they are either causal or non-necessitating. I think whenever we feel that reason dictates something this is not a case of necessitation in the real hard sense. The "dictating" in such cases is done by a very effective norm, which cannot be breached without completely messing up one's mental life. Nevertheless, it is just a norm, and breaching it is not a complete impossibility in the literal sense. We normally feel that we could never think that 2 by 2 is 5 because we are normally completely unmotivated to think such a silly thing. Were we motivated strongly enough, we could do it. Presumably, our minds would have to have already been messed up to a considerable extent for being so motivated. Empirically, there are cases when people refuse to acknowledge things whose obviousness is comparable to that of "2 by 2 is 4". So I am now considering an option that I think doesn't really exist: when the relation between the reason and what it dictates is really mechanical, although the mechanism is not causal. Can I think of a real mechanism that is not causal? No. But maybe someone else can. I am discussing this option for his sake.

morally relevant sense. We have seen earlier in this chapter that events indeterministically caused by reasons, with the assumption that nothing else apart from their indeterministic causes is relevant to the question whether they occur or not, do not constitute activity in the required sense either. The conclusion that we are facing now is that explicability in the light of reason, even if reasons are thought to work non-causally, cannot be either what activity consists in.

Perhaps it is time that we draw the conclusion that activity is not constituted by a choice's relation to the chooser's reasons.

If it is a reality, it must consist in something else.

Strawson rightly thinks that at this point the libertarian has to posit a power to determine ourselves that is *not derivative of our capacity to be rational*. But he wrongly infers that if genuine freedom of choice does not consist in a relation between reason and choice then it is a mere non-rational flip-flop.

Here is what a libertarian can suggest:

Activity in respect of a choice consists in the choice's being related to the agent, rather than to his reasons, or to anything characteristic of what he is and how he is at the time of the choice, in a certain way, i.e. that the agent bears a conceptually primitive creative relation to the choice.

I know this suggestion is not easy to swallow. It is positing a mystery. It is part of the philosophical price we have to pay for libertarian freedom. At the moment I don't want to argue about the bearability of this price, or about how it relates to the price we would have to pay for abandoning libertarian freedom. In this chapter my purpose is to argue only two things (apart from what I have already argued, i.e. that this is the only way to save libertarian freedom). One is that there is nothing incoherent about this suggestion. The other is that it doesn't reduce genuine freedom to irrationality. We are at the moment particularly concerned with the second of the two.

Surely this is a power that can be used to choose irrationally. It is the power to act in an *arbitrary* way. This arbitrariness is not to be confused with randomness. It is actually in virtue of this arbitrariness that choices underdetermined by causes and unnecessitated by reasons can be non-random.[293]

---

[293] My usage of the adjective "arbitrary" may be unusual, or even completely alien to how it is used in modern English. Not being a native speaker of English, I cannot really tell. I was advised that "arbitrary" is largely considered synonymous with "random". So please take it as a technical usage, which nevertheless has something to do with the non-

There are cases when this power is used in a way that there is no nontrivial answer to the question why the agent did what he did.

Yet, this is also the power to act rationally. Not only in the sense that this power can also be used to make choices that are reasonable, but also in the sense that we can choose, or infer, or make a judgement rationally *only in virtue of this power*.

What led to Strawson's conclusion that this power is a non-rational flip-flop was that we tried to distinguish some causally undetermined events from random events with reference to the work reasons do in their production, and we failed. But "the work reasons do" in the production of an action (a choice, a judgement) is just a misleading *façon de parler*. Reasons do not work in producing the action (the choice, the judgement). Agents do. The work can partly be described as considering reasons and weighing them against each other, this part of the work may be called "reasoning", but *the result is not attributable to a power or potential inherent in the reasons*. The work is terminated by an exercise of this power to create by the agent. If my arguments in chapter 5 are sound, then without the exercise of this power there would be no rationality, either in practical, or in theoretical contexts.

So rather than being an "entirely nonrational flip-flop", this power is what rationality rests on.

In chapter 5 we came to the conclusion that the Epicurean intuition that libertarian freedom is a prerequisite for rationality was correct. Epicurus thought that if thoughts are produced by causes, be them causal powers inherent in reasons, then they are produced the wrong way, and they cannot be rational. Because rationality is activity—evaluating, judging, and committing ourselves to

---

technical usage. It draws on the Latin origin of the word. My Latin is much worse than my English, more precisely, my Latin is practically nonexistent, yet, I know that much that *arbitrio* means "will". So my "arbitrary" could translate as "just willed". But saying "arbitrary" instead expresses more. I say "the primitive power to create" is the power to act "arbitrarily", because I want to emphasize that it is a power to bring about uncaused events and it is also a power with which we can breach the norms of rationality occasionally—although it is also a power to follow the norms of rationality, if we hadn't this power, the laws of rationality would have to be descriptive laws rather than norms, which they aren't. Yet our "arbitrary" actions are not random, precisely because they were determined by us, we determined them simply by performing them, using our primitive power to create. In contrast, my paradigm of randomness rather that "arbitrariness" would be a causally underdetermined quantum mechanical event, like a decay of a heavy atom, which was not done by anybody.

propositional attitudes or courses of action, and causal determinatedness, in turn, is passivity. If there is a full causal explanation for our commitments to thoughts or courses of action, then there is no room for a truthful rational explanation for the same commitments. A rational explanation for them would be a mere *post facto* rationalization, not telling the truth about why we committed ourselves to these thoughts or courses of action. The reasons cited in those explanations would be epiphenomenal. I argued that the Davidsonian suggestion that the two explanations might be one and the same, because reasons are causes, either fails to account for the anomalism of the mental and the normative (in contrast to descriptive) nature of the laws of rationality, or accounts for these phenomena at the cost of falling back to epiphenomenalism in respect of reasons (their mental content).

Davidson thought that reasons needed to be causes because he thought that if there weren't a causal relation posited between the reason for which the agent acts and his action, then something important would be missing from rational explanation.

He writes that if rational explanation is no more than just an appeal to

> certain beliefs and attitudes in the light of which the action is reasonable...then something essential is left out, for a person can have a reason for an action, and perform the action, and yet this reason not be the reason why he did it. Central to the relation between a reason and an action it explains is the idea that the agent performed the action *because* he had the reason.[294]

So far we agree. That is the ground on which, following Epicurus and Lewis, I argued in the fifth chapter—in the context of theoretical rather than practical reasoning, but the idea is the same in both contexts—that a rational explanation cannot be truthful about why what it rationalizes took place with there being a parallel independent causal explanation that fully explains it.

But Davidson goes on to claim that unless this "because" that links the reason to the action is construed as a causal relation, the nature of this relation remains a complete mystery.[295]

---

[294] Davidson 1963, p. 9. Stress in the original.

[295] Ibid, p.11.

Why would that be so, I don't see, and Davidson doesn't tell. There is a perfectly cogent account of the "because" that isn't necessarily causal. I suggest that what that "because" means is something along the following lines. The reason "because of which" the agent acted was not just a combination of a belief and a desire that the agent happened to have, it has actually occurred to the agent that the course of action that is being explained he believes to fulfil the desire; and the agent actually *committed himself to try and fulfil that desire by performing the action*.

The commitment perhaps followed an evaluation process during which it occurred to the agent that he believes that if he chooses this course of action some other desires he also has will be frustrated. He weighed the incompatible desires against each other and judged that he values more the one that would be fulfilled than those that would be frustrated. Maybe he has more than one desire that he believed his action would fulfil. Maybe he chose the way he did because he judged that he values more these desires taken together than those that would be frustrated by the same course of action taken together. If he did so, then there is no point in asking which particular belief-desire pair was it for which he chose the way he did. It was all of those that comprised of a desire that occurred to him combined with a belief that his effort to fulfil them by the course of action he chose will be successful. And it is also possible that he saw very clearly that if he chooses this way some desires he values more, e.g. being the kind of man who doesn't behave so in the given type of circumstances, will be frustrated, and he committed himself to perform the action anyway. It is also possible that he didn't have the time or the patience to sort out his conflicting desires properly, and made up his mind to fulfil the desire in question even though he wasn't sure he values it more than the conflicting ones.

The details of the story could be filled in any of the ways just sketched. What seems to be essential is that both the desire and the belief that the course of action will fulfil the desire must have occurred to the agent, and he must have committed himself to fulfil the desire by performing the course of action, if the reason that is the combination of the desire and the belief is to figure in a truthful answer to the question why the agent did what he did.

I don't see why this account, with the details filled in any of the above sketched ways, would need to be a causal one. And I don't see that anything essential would be missing from it.[296]

The relation between the reason and what it truthfully rationalizes (be it a thought or a practical choice) doesn't need to be a determining one in order to secure the rationality of the latter. On the contrary, if what was argued in chapter 5 is correct, then it shouldn't be. The picture that (practical) rationality is essentially the exercise of "the primitive power to create" by performing an undetermined action in order to realize a desire that the action is believed to realize seems perfectly cogent to me. It is also very much in line with our experience that we are rational but not perfectly rational, that we act in loose conformity with reason, with the laws of rationality being norms that we either follow or don't follow, rather than descriptive laws analogous to the laws of nature. Our actions' loose conformity with reason can be seen as some, though weak and indirect, empirical evidence in support of this picture.

The coherence of the conception of the primitive power to create

We have now clarified the relation between rationality and the "primitive power to create" and found that what we do undeterminedly by either causes or reasons is not bound to be non-rational, on the contrary, it is in virtue of this power that some of our actions are rational. Now it is time to address the question whether the conception can be coherent.

I said that this power is the power to act arbitrarily in the sense that it is uncaused and unnecessitated by reason. Can such an arbitrary event be non-random?

There are philosophers who say it is easy. Carl Ginet, for one, says it is non-random because the agent determines it. If an event is causally undetermined, then it is impossible for the agent to bring

---

[296] In an interesting article I have already cited in the fifth chapter Julia Tanney (1995) draws on a note by Davidson in support of his thesis that reasons are causes saying "A desire and a belief of the right sort may explain an action but not necessarily. A man might have good reasons for killing his father, and he might do it, and yet the reasons not be his reasons for doing it (think of Oedipus)." (Davidson 1974, p. 232). Tanney, thinking of Oedipus; considers different cases Davidson might had in mind in relation to Sophocles's story in which it may be *prima facie* problematic to discern the exact reason on which Oedipus acted, and gives a non-causal account of each situation that identifies the reason that Oedipus not just had but acted upon.

about causal conditions that would necessitate its occurrence. But in case the causally undetermined event is the agent's own action, then determining that it would occur requires nothing like that.

> [I]t requires only that one perform it; and one performing it, which is just the action's occurring, is compatible with the action's being undetermined, not causally necessitated by antecedents.[297]

But how is it different from random exertions of the body? An involuntary blinking of my eye would be a good example. In a technical sense, it is something I do: neurons fire, muscles move, all that taking place in me. But Ginet would not want to say that I determined that the blinking would occur by doing it. That would invoke a different sense of "doing". How are the two senses different?

Ginet answers this question by refining his account. An exertion of the body, if it is an action, is a complex action. The agent does not determine it simply by doing it. That is true only of simple actions. At the core of every complex action there is a simple action, which the agent determines by performing it, and the rest of the complex action is causally necessitated by the simple action. Simple actions are volitions. They are mental acts that differ from passive mental occurrences in virtue of an "actish phenomenal quality", which they uniquely possess. They have an intentional content which is directed at the immediate present, so they are temporally co-existent with the whole complex action.[298] So it is the simple mental action at the core, possessing the actish phenomenal quality that distinguishes the "true" sense of doing from the merely technical one.

Well, for all I know introspectively about the genealogy of my actions, any mental occurrence with an "actish phenomenal quality" can be a genuine mental action of which I am active in the sense required, but it can as well be a random occurrence, or can even have a necessitating cause. There is nothing in the introspective phenomenology of action-production that would warrant me against that. What Dennett points out in *Elbow Room* in respect of decision making in general is true of the forming of volitions directed to the immediate present, too: "We have to wait and see how we are going

---

[297] Ginet 1990, p. 127.
[298] Ibid. pp. 12-32.

to decide something, and when we do decide, our decision bubbles up to consciousness from we not know where. We do not witness it being *made*; we witness its *arrival*"[299] The way volitions are invoked in Ginet's account I find adequate in other respects, but their having an actish phenomenal quality does not guarantee that the complex action of which they are the core is after all not produced randomly (or deterministically). The distinction we are after, i.e. the one between arbitrariness and randomness, cannot be based on this ground.

But it doesn't need to be grounded in such a way. Ginet seems to be looking for an intrinsic quality that would distinguish between an event which is an unnecessitated simple action of an agent and an event that happens at random to the same agent. But maybe there is no such intrinsic quality, and maybe such an intrinsic quality is not necessary to distinguish between the two. Maybe two mental events may share all their intrinsic qualities, yet one may be a genuine volition and the other just a random mental event, and their relational qualities distinguish between them.

But there is no way to distinguish between the relatedness of the two to prior events or states of affairs in a language that doesn't make an irreducible reference to the agent. The distinction cannot be accounted for in an impersonal language.

Suppose that the agent has two conflicting sets of rational considerations R_A and R_B, one supporting action A, the other supporting an incompatible action B. Suppose that R_A and R_B are quite evenly balanced, so even after a long series of efforts to choose between A and B the agent is undecided. Now he suddenly decides that he would do A for R_A and does so. How is it different from another case when his will to do A for R_A emerges out of the intricate, maybe at bottom quantum mechanical, processes of his mental life just at random?

The only way to account for the difference in an impersonal language would be to posit a causal relation between R_A and A, as Davidson does. But if we do so, we forfeit freedom and also rationality, as we have seen. The only remaining option is to accept that *there are facts that cannot be accounted for in an impersonal language.*

The account for the difference between the two cases in a personal language would be something like that in the first case the agent formed the volition to do A by using his power to make up his

---

[299] Dennett 1984a, p. 78.

mind, while in the second case he didn't. And there is nothing more to it.

I cannot give a further account of what "the power to make up his mind" would mean. If it is anything, it is a power that cannot be reduced to the causal powers inherent in any property that can be truly predicated of the agent in an impersonal language at the time he makes up his mind (whether or not the impersonal language is allowed to contain items referring to irreducibly mental objects or properties).

I agree that it is puzzling. The ontology of persons must be really weird, quite different from the ontology of ordinary objects.

I have already acknowledged that I don't have an explanation for that power. But it doesn't mean that it is incoherent—it may be just conceptually primitive.

There is a lot more philosophical work to do here to develop the ontology of persons having the primitive power to create. But I don't see why would the idea of a person having a power that is not reducible to the causal powers inherent in the properties that can be truly predicated of him in an impersonal language bound to be incoherent, and we have seen that identifying genuine freedom with cases when that power is exercised doesn't reduce freedom to irrationality.[300]

---

[300] Am I advocating an agent causal theory of action, after all? Not exactly. My view can be seen as being halfway between agent causalism and Carl Ginet's position, which was dubbed "simple indeterminism". I agree with the agent causalists criticizing Ginet that the property that distinguishes genuine actions from random occurrences is not an intrinsic property, inherent in the events that are genuine actions, but a relational one, characteristic of the relation that holds between the agent and his actions. Yet, I don't think this relation should be construed as a species of the causal genus. I think all cases of event causation we have ever empirically learned of rest on causal powers inherent in properties (universals) that the cause instantiates to bring about the effect. Known cases of event causation are a regularly occurring relations between instantiations of universals. Otherwise we wouldn't know about them. There is nothing like that in the case of the relation between an agent and his action. An action is a free-floater in respect of all properties the agent instantiates at the time he does it. If it were not, then the action would not be free. Unless of course actions are attributable to properties with powers that we are capable of mobilizing, or keeping at bay. But then the same question would arise concerning the event of mobilizing them. If to this question the answer would be that the question is wrongheaded because the mobilizing is not an extra event over and above the agent's performing the action that should be explained, that is very fine, but then how is the property to which the single event to explain was attributed explanatory of the event represented or described as the event of mobilizing the power the property conveys? I think it is much better to account for the non-randomness of genuine action the way Ginet does. He says an action is non-random in virtue of being determined, and

Answers to the problems posed at the end of chapter 4, a further question, and some closing remarks

At the end of chapter 4, in which I argued that our moral intuitions endorse a U-condition for moral responsibility, three questions were left hanging in the air concerning the U-condition.

The first two were whether the U-condition was coherent and whether by requiring the U-condition for responsibility we are requiring something really absurd, namely, that in order for someone to be responsible for his deed, the deed must be an irrational one.

In this chapter I have said everything I have to say to answer these questions. I hope the answers were satisfactory.

Before we attend to the third question raised in the end of chapter 4, I would like to address a further question. Throughout this chapter whenever an attempt to distinguish some underdetermined events from mere randomness in view of identifying cases of genuine activity was considered we enquired about whether the proposed distinction was relevant morally, whether it was reasonable to think that the agent can be morally responsible for the proposed cases of activity more than he is responsible for a genuinely random blinking of his eyes for example. In the proposals discussed earlier the suggested cases of genuine activity were related to the agent's reasons some way or another, and that was supposed to be constitutive of the kind of events for which the agent is responsible. On the account I am proposing there is no relation between reason and action that has to obtain for an action to be a genuine one. On my account genuine actions are genuine actions in virtue of an arbitrary element, an exercise of a power which I dubbed the primitive power to create, of

---

it is determined by the agent, and the agent determines it simply by performing it. And there is nothing more to be said about it. The relation between the agent and the action is probably primitive both conceptually and ontologically. If we accept Ginet's view that at the core of every action there is a simple action, which is a volition, then it may be unusual to talk of agents performing volitions, but I see nothing wrong with it apart from the verbal unusuality. Whereas, as far as I can tell, agent causalists are mainly concerned with how the relation between the agent and the action should be made intelligible as a causal relation and how it relates to event causation, I think the interesting philosophical question is, rather, how we should make intelligible an entity—a person—who is capable of transcending himself, in the sense that he has the power to create a new edition of himself that cannot be derived causally from the old edition that he is at the time when he endeavours the creating. I plan to address this question on another occasion.

which all that there is to be said is that it occurred, although it might not have. Are genuine actions then distinguished from random occurrences in a morally relevant way? Why is it so obvious that we are morally responsible for them?

I think this question is motivated by the worry that on the account I am advocating we are deprived of all means to control our actions. Neither reasons nor any other mental object, event, or state of affairs can be identified as the means through which we control that arbitrary element which, I suggest, has to be at the core of all cases of genuine activity. The answer to this worry is that the suggestion is precisely that no such means are needed. Genuine action is controlled by nothing *in* us, but *by* us. We control it by performing it. This is the only truly responsibility-conveying sense of control.

Now that it has been clarified, the third question raised in the end of chapter 4 can be properly addressed. It was essentially the question that, provided that there are cases when we are really morally responsible, can those cases be reliably detected. Can cases when the U-condition is met be discerned reliably? Because if not, then the U-condition has no practical applicability in our ordinary moral practices.

I have to admit that I find this worry well grounded. I seem to use my power to create, to perform self-forming actions in Kane's sense that meet the U-condition, quite often. But I may be wrong. I have admitted to Dennett that, for all I can know for sure introspectively, I could be a robot. Or a robot equipped with a randomizing device.

And of course I know much less of others than I know of myself.

All I think I can argue for is that there is no good reason to think that we are robots, randomized or not. I think usually it is a good philosophical policy to hold on to common sense as long as we are not forced to abandon it by convincing arguments. I do seem to exercise my power to form volitions to perform actions that I believe will lead to the satisfaction of my desires, and I do seem to have the power to repent from this, and it doesn't seem to be just a brute random fact which way I choose. This is how we commonsensically think of ourselves, and anyone wanting to talk us out of these commonsensical beliefs about ourselves should offer good arguments. Throughout this thesis I have shown the inconclusiveness of several arguments purported to show that this cannot be true.

On the other hand, it is also part of common sense that we feel to be moved by desires. I often seem to feel the force exerted on me by

them. To what extent I am moved by them helplessly, without my will to interfere to alter the course of events, I cannot know. It doesn't even seem to be an all-or-none issue. Some forces moving me seem stronger than others, and harder to resist.

As far as blaming is concerned, giving in to very strong, though in principle resistible, nonagreeable motifs, in respect of which an agent is passive, intuitively invites easier treatment than when the forces that move us are weaker. There seems to be a full spectrum of corresponding different degrees of guilt stretching from full responsibility to complete exemption.

But these considerations are not very helpful if we have to decide in concrete situations whether we should blame someone or not.

My attitude toward this question is that perhaps it is best to distinguish between public and private responsibility-attributions, and between the institutionalized retributive practices of the society on the one hand, and ascriptions of moral responsibility that is meant to be metaphysically adequate, on the other.

As far as the former is concerned, having regard to the obvious social utility of holding people responsible, and the hazards of abstaining from it, it may be justifiable to hold people generally responsible for what they seem to do voluntarily by default, and look for exculpating circumstances on the basis of the U-condition. In one perfectly good sense it would be morally wrong to abolish our public retributive practices, even if we are aware of the fact that they are fallible.

As far as the latter is concerned I am inclined to suggest that we should hold back from blaming people as much as we can. There is no obvious harm associated with our private reluctance to hold people morally guilty on the basis of our limited epistemic access to the facts that determine whether their actions were U-condition satisfying, or not. So in this case holding back from blaming seems to be the morally right thing to do. As put by Cornelius Plantinga,

> Cultural influences, personal strengths and insights, the human capacity for self-deception, conscience as shaped by "the law of God written on the human heart," and numerous other factors combine in such intricate ways that we are seldom in position to make accurate judgments about even our own blameworthiness, let alone someone else's. Judgments about degrees of culpability, unless

> required by such special roles as parent, judge, or jury, may therefore wisely be left in the hands of God.[301]

I know, in some respect, this is a disappointing answer. But I think it is an important philosophical conclusion that we may have objective responsibility for what we do, even though we can never be sure whether someone is objectively responsible in a particular case.

[301] 1993, p. 190.

9 Conclusions: The Philosophical Cost and Benefit of Libertarianism

The two problematic alternatives of thinking about alternatives and control

Our discussion started with recording that on our intuitive conception of freedom we are free in situations where both of the following two statements are true: 1) Given everything that has already been laid down (the past and the present) there is a multiplicity of objective possibilities for the history of the world to continue. 2) We have the capacity to control which of these possibilities will occur by choosing how we act. In short, we are free if we have both alternatives and control.

These two statements can hold true simultaneously only if there is a way of making sense of control other than construing it as a special case of causal necessitation of what is to come by what has already been laid down.

So one way of thinking about freedom is to claim that freedom consists of the simultaneous truth of 1 and 2, supplemented with an account how control is achieved, given that it cannot be achieved by way of causal necessitation of what is to come by what has already been laid down. This is the libertarian way of thinking about freedom.

The other way is thinking that control is a species of the causal necessitation of what is to come by what has already taken place. If one thinks this way of freedom, one has to distinguish between different cases of causal necessitation of what is to come by what has already taken place on the ground of some principle. If the causal chain leading to one's action is of some specific sort, if the immediate causes of action are internal to the agent, such that the agent can identify with them, or the agent can reasonably be identified with them by the relevant moral community, then, causal theorists tend to claim, the causal necessitation of action by what has already taken place is not incompatible with freedom. Otherwise it is. This is the causal way of thinking about freedom. This is compatible with determinism, so it is also called the compatibilist way.

Both ways have their problems.

The problems with the libertarian way are these: The libertarian must supply an account of how control can be achieved without the

causal necessitation of how we act, or how we make up our mind to act, by what we are and how we are at the time. It is not obvious that it is possible. He also has to show that control can be rational even though on his account reasons are not allowed to necessitate free choices. It is also worrying that many scientists and philosophers hold that it is never the case that there are objective possibilities in the future, either because of scientific results that point to determinism, or for some concerns about time, like the relativity of simultaneity, which seems to undermine the ontological difference between what is yet to come and what has already become real, indicating that there is no such thing as an open future.

The problems with the causal way are these: Although some compatibilist philosophers made efforts to prove to the contrary, if control is construed as a special case of causal necessitation of what is to come by what has already been laid down then there are no objective alternatives. There may be alternatives in the subjective sense, meaning that the causal chain that leads to a choice may be such that putative alternatives are considered, weighted and chosen from, all in a deterministic way, and this is how the only objectively possible outcome is brought about. Consequently, control cannot be understood as controlling which alternative is to occur ("regulative control"), it can only mean that what we are and how we are has a necessary role in the causal production of what we do ("guidance control"). It is far from obvious that the existence of alternatives in the subjective perspective and control in the sense of guidance control are sufficient to ground the values that we normally associate with freedom, such as self-determination, moral responsibility, rationality and intellectual responsibility. It is also a problem that the causal conception of freedom is necessarily shallow or superficial in the sense that the causal theorist can posit requirements which, if met, guarantee that causes may necessitate our choices without diminishing freedom only in respect of a finite number of links going backwards along the causal chains that lead to our choices.

Whose problems can be solved?

In this thesis I argued for the following major claims.

1) Despite the ingenuity of some compatibilist philosophers who tried to prove to the contrary, *if we adopt the causal conception of control we do forfeit genuine alternatives* (chapter 2). I argued that the "inability

operator" in the consequence argument understood in Timothy O'Connor's strong sense is universally closed under conjunction and logical entailment, so the consequence argument is valid. I also argued, in particular against David Lewis's "local miracle compatibilism", and against "multiple pasts compatibilism" advocated by John Turk Saunders and others, that the premises of the consequence argument hold, unless we assume, like Kant, that the flow of time is only phenomenal, posit a timeless reality, and assume that freedom is exercised timelessly. On these assumptions the premise that the past is not in our power to make different can be challenged. Freedom and determinism can be made compatible this way, but this Kantian freedom has nothing to do with the causal conception of freedom, it is the libertarian freedom of the timeless "noumenal" self.

2) *The shallowness of control, on the causal conception of it, is a serious concern for self-determination* (chapter 3). I argued, against Daniel Dennett, that this shallowness is not necessary for self-determination to be practical, and that the libertarian conception of self-determination does not require us to create ourselves out of nothing. I argued, following Robert Kane, that this shallowness makes agents who are perfectly free on the causal conception of freedom possible victims of "covert non-constraining manipulation". I argued against the anticipated Dennettian objection that such situations, like the Walden Two thought experiment used by Kane, would be "unfair intuition pumps".

3) *The senses of moral responsibility that are available without genuine alternatives leave our intuitions about what moral responsibility requires unsatisfied* (chapter 4). I argued against Dennett's direct arguments, dependent or independent on his substantive evolutionary theory of morality, to the effect that it is fair to hold deterministic agents responsible. I argued against him and Harry Frankfurt, following Peter van Inwagen and Martha Klein, that our moral intuitions support a could-have-done-otherwise condition for moral responsibility, and I also argued that it should be understood in the objective sense, not in some subjective sense, or in the sense of Hume's "hypothetical liberty", or G. E. Moore's conditional analysis, or Dennett's "personal stance". I added to these arguments in chapter 8 by arguing that even Humean compatibilists, like Robert Adams, who are committed to the view that it is fair to hold people accountable for the deeds that proceed from their character a

necessary way, because these are informative about their character, and it is fair to blame people for their character they cannot help to have, should accept the principle that no one is more responsible for a response one's character produces necessarily to a stimulus than one is responsible for one's character, provided that one is not responsible for the stimulus (TP2). I argued that once this principle is accepted there is no way to deny the principle that moral non-responsibility is transferred through causal necessitation (TP1'), and that this principle is incompatible with holding deterministic agents responsible.

4) *Rationality and intellectual responsibility cannot be achieved by a mechanism* (chapter 5). I argued that the Epicurean intuition that we cannot properly be said to possess rationality and intellectual responsibility unless we are free in the libertarian sense was correct. The Epicurean argument rested on the intuition that if there is a causal explanation for why a thought or an intention to act arose, then a rational explanation for the same can only be a mere *post facto* rationalization that does not tell the truth about how and why the thought or the intention came about. I argued against the Anscombian objection against this intuition that the causal and the rational explanation of the same mental event can be conceived as two independent matters that do not compete with each other. I also argued against the Davidsonian objection that reasons may be causes, so the causal and the rational explanation may account for the same genealogy of a mental event under two different descriptions. As long as the anomalism of the mental is accepted as an empirical fact, and causation is assumed to be nomological (which are the assumptions motivating Davidson's anomalous monism), both objections to the Epicurean argument lead to the view that reasons *qua* reason are epiphenomenal, and whether what we are caused to think is also rational to think, is a mere matter of luck. I argued that no evolutionary argument can amend this situation.

5) *As far as our present knowledge goes, determinism is empirically unfounded* (chapter 6). I argued that psychological determinism is far from being an empirical fact (as for example Hume famously claimed it was). I also argued that even if a deterministic interpretation comes out winning from the present debate about how quantum mechanics should be interpreted, of which there is no clear positive indication at the moment (as it can be seen in the Appendix), physical determinists should also prove that the laws describing the physical evolution of

systems containing conscious minds, or with which conscious minds interfere, are also deterministic, and that nothing like this have so far been proved, or even made plausible. In response to Ted Honderich's claims that neural determinism, which is the determinism of a system which trivially interferes with a conscious mind, is unanimously accepted among neuroscientists, so quantum indeterminacy, if there is such a thing, is irrelevant—given that I am not competent in neuroscience—I could only cite the equally assured testimony of other neuroscientifically informed philosophers, Henry Stapp and John Eccles, to the contrary. Both of these philosophers pointed to ways quantum indeterminacies could propagate to the macro level in the evolution of brain states. I argued that those philosophers, e.g. David Papineau, who think that Eccles's suggestion that the mind could control its brain by "biasing the Born rule" in quantum mechanical processes at the synapses rests on an elementary mistake because it would contradict the probabilistic laws of quantum mechanics, are in an elementary mistake concerning the nature of probabilistic laws on both the frequency account and the propensity account of probability.

6) *As far as our present knowledge goes, the future may well be open* (chapter 7). The argument to the effect that there is no open future I was concerned with was the argument from the special relativistic symmetry of inertial observers offered by Kurt Gödel, Hilary Putnam and others. I presented several ways the thesis of the relativity of simultaneity could be resisted consistently with the empirical data that led to the special theory of relativity, including arguments from physical cosmology and from quantum non-locality for absolute simultaneity. I argued, however, along the lines of Howard Stein and Dennis Dieks, that we do not have to rely on these arguments, because we can both embrace the relativity of simultaneity and hold on to objective becoming and the ontological openness of the future, if we conceive of the present locally. I argued against Simon Saunders's objections to Stein's and Dieks's local presentism that it would breach the relativistically meaningful requirement of intersubjectivity with respect to temporal determinations that are meant to have an ontological significance, and that it would lead to solipsism, and I hope to have refuted both.

7) *The libertarian conception of control involves no contradiction, and it can be rational, in fact, only libertarian control can be rational* (chapter 8, relying on chapter 5). I argued that the "causal indeterminism" of Robert Kane

and Robert Nozick fails to distinguish libertarian free choices from merely random occurrences in a morally relevant way. I also argued, improving on a regress argument of Galen Strawson, that David Wiggins's suggestion that some undetermined choices are non-random *because* they are intelligible in the light of the agent's reasons, cannot be the solution to the coherence problem of libertarianism. I argued that the only way to make sense of libertarian control is to posit a conceptually primitive relation between agents and some events that is the relation of a creator and its creature, and endow agents with a "primitive power to create", which cannot be traced back to any causal power they have in virtue of properties that can be truly predicated of them (facts that hold true of them). I argued against Strawson that this "primitive power to create" is not a mere "non-rational flip-flop of the soul", quite to the contrary, having regard to the findings of chapter 5, this is the capacity in virtue of which we can be rational. I also argued that the idea of this power is coherent, even though it requires a very radical non-reductionism about persons.

So, crudely put, I argued that the problems with the causal (compatibilist) conception of freedom are insurmountable, whereas the problems with the libertarian conception of freedom aren't.

Another way of putting it would be to say that the philosophical benefit of libertarianism is well worth the philosophical price we have to pay for it.

The philosophical price of libertarian freedom

It is certainly part of the philosophical price of libertarianism that it embraces a concept of freedom which may turn out to correspond to nothing in realty if the further advancement of science proves that the evolution of our psychological states is deterministic or that there are no future contingencies.

In relation to this I argued for two claims.

One is that this possibility is much more remote than some philosophers (e.g. Fischer and Honderich) think it is, which would amount to saying that this price is not very big.

The other is that from this possibility we should not derive the philosophical policy that we should develop our metaphysical theory of freedom in such a way that it would not be necessary to react to the discovery that determinism is true or that there is no open future

with the horror of the loss of freedom. I argued that the philosopher's work is to sort out the freedom-related values that would be lost and those that could be retained if such a discovery was made, which is to say that this price should not be taken into account as a principle of theory choice in the metaphysics of freedom, as, for example, Fischer seems to have suggested.

Surely the toughest of the problems for libertarianism is the problem of the coherence of the libertarian conception of freedom, i.e. whether we can make sense of control over which of the alternatives will take place in such a way that does not require that the alternative that takes place should be causally determined by what we are and how we are at the time (given the circumstances).

There is nothing mind-blowing in the suggestion that the correct analysis of randomness is not that randomness is the same as underdeterminedness, but that an event is random if and only if it is underdetermined and it is not an action. There is nothing obviously incoherent in this proposal. We have no obvious a priori knowledge of the nonexistence of a third category of events besides determined and random.

But of course for this suggestion to be taken seriously the libertarian theorist has to explain how the events falling in this third category are non-random. We understand how they differ from determined events: they don't have the sufficient causal conditions for their occurrence in the complete previous history of the universe. But how are they then different from events that pop up just at random? And the libertarian cannot just point at any difference. The difference must be relevant to the values that the libertarian wants to save, in respect of which he finds compatibilist freedom disappointing: self-determination, moral and intellectual responsibility. In the eighth chapter we have reviewed some libertarian strategies to distinguish a third category of events from random events that fail at this point.

The problem can be viewed liked this.

In order that there be action and not only passion, there must be choices of which it is true that they are controlled by the agent, but it is also true that they are not determined by any fact—be it a fact about the agent or of the rest of the world—that obtains prior to the making of the choice. So the agent cannot control such a choice in virtue of any fact that obtains in him, in virtue of any fact the

obtaining of which is constituent of what he is, who he is, or how he is, mentally or physically speaking.

Therefore, if he is to control such an undetermined choice, there must be more to him than the totality of the facts about him, mental or physical, and he must have other means of control than the causal significance of the obtaining of all these facts.

So, if we want to believe in libertarian freedom, we have to adopt a radically nonreductive view of personhood.

We have to believe that there is some ontological primitiveness to personhood, in the sense that persons are not constituted by facts, at least not fully.

If we swallowed this, we might suppose that the person, who is more than the sum of all facts about him, has a power to create. This is a power to create events, the obtaining of new facts, in a sense *ex nihilo*, that is, to bring about facts that are not necessitated by previously existing facts, the obtaining of which, then, can be the starting events of causal chains that otherwise would have never existed.

It is a godlike power, the power of a prime mover, who is not moved by anything. Yet, this mover gets moved. The facts that he brings about change him. By bringing about new facts the person brings about new editions of himself, too. In this he is moved by himself, but himself understood primitively, i.e. not by anything, or by the totality, of what was true of him previously.

This power to create must be conceptually primitive. It is a relation that holds between a person and an event, and there is no further story to tell about it. If there was such a story to tell about this relation, then either the event was not the immediate beginning of a causal chain, or we would invoke facts about the person in virtue of which the relation between him and the event holds, and then we would be back with the dilemma of sufficient or insufficient explanation in terms of facts, and that would lead back to the dichotomy of determinism or randomness.

The philosophical price we have to pay for libertarian freedom is the positing of these ontologically curious entities, persons unconstituted by facts. This price is clearly unbearable for materialists, and pose serious difficulties for dualists, too.

But the philosophical price of the alternative, giving up on libertarian freedom and embracing the causal conception of freedom instead, is even heavier.

The philosophical benefit of libertarianism

The philosophical benefit of libertarianism is that we don't have to pay the philosophical price of its alternative.

Compatibilists argue that this price can be taken lightly.

I admitted at several points of the discussion that the distinction between different ways the causal production of action may take place, on which the causal conception of freedom relies, is not completely irrelevant for freedom. I acknowledged, for example, that it is a worthwhile sense of self-determination if our life unfolds from what we are, from the desires and the values we embrace, and what we hold true of the world, and if our actions arise from this through a deliberative process in which we canvas and evaluate alternatives that seem possible to us if we want them, in contrast to the situation when our actions arise from coercion, or from a disorder of our deliberative faculty, and our life is dominated by oppressive environmental influences that override our desires and values. This sense of self-determination is available to us on the causal theory, if we are lucky. So, in this sense, there isn't only one legitimate way of conceptualizing about freedom.

Ted Honderich argues that both compatibilism and incompatibilism rest on a mistake, and the same one, namely, the conviction that there is a single correct analysis of the term "freedom". Both of these big traditions are fixed on their holy grail, the true analysis, both of them tragically believe themselves to be in possession of it, while, of course, it is not to be found. This is the main reason why there is so little progress in this subject of philosophy. The metaphysics of freedom, before Honderich, who finally made the discovery about the plurality of respectable analyses, was a dialogue of the deaf. Or so he is claiming.[302]

The alternative he is offering to the misconceived dichotomy of compatibilism and incompatibilism he calls *attitudinism*. To attitudinism one arrives by realizing the plurality of different important senses of freedom, and by realizing that they are associated with different contents of some normative attitudes. Honderich thinks, and so do I, that these attitudes have different contents depending on the sense of freedom that is tacitly presumed in them.

---

[302] 2002, chapter 9.

Some of these are meaningful on the causal conception of freedom, some only on the libertarian conception.

He asks how we should react to the "near fact" that determinism is true, and so the causal conception of freedom is the only live alternative. His answer is this: If we strongly attach ourselves to the attitudes that are dependent on the sense of freedom which is incompatible with determinism, we will react with the feeling that the world is regrettably a much duller place than we previously thought it was. But there are those attitudes whose content is not dependent on that sense of freedom. Why don't we embrace them and let go of the rest? These are just attitudes, that is, normative statements, and, "as we all know", normative statements are neither true or false. This embracing of some attitudes and letting go of others Honderich calls *affirmation*. Since there is no truth about norms, he says, we can perform affirmation without breaching truth or intellectual honesty.

This suggests that the philosophical price of choosing the causal theory of freedom is soft. It depends on our attitudes towards it. It can be taken lightly.

I don't think it is so simple. It is meaningful to argue about many normative matters. For example, it is meaningful to argue about whether it is more *desirable* to be the captains of our fate in the sense that involves a plurality of options even relative to what we are and how we are at the time, than being the captains of our fate merely in the sense that not someone else is the captain. Similarly, I think it is meaningful to argue about whether it is more *valuable* to be loved with a love whose genealogy involves a self-forming decision from the part of the one who loves us, than with a love that is merely free from external coercion. I also think it is meaningful to argue about whether it is *fair* to hold deterministic wrongdoers blameworthy. These are all normative arguments. It is true that arguing about such matters one usually ends up invoking unargued, intuitive convictions. But that is quite common in philosophy, certainly not unique to normative questions.

It is bad if we can only have a kind of self-determination that is shallow in the sense that has been discussed; if moral responsibility reduces to the regrettability of a bad character (Hume, Adams), or the evolutionary adaptiveness of responsibility-attributing practices (Dennett), or to the fact that with our social behaviour we tacitly express that we agree to be held responsible if our actions are causally produced by an internal mechanism that meets some requirements

(Fischer and Ravizza); and if rationality is reduced to the smartness of an algorithm, and there is no way to make sense of intellectual responsibility.

What is at stake here is what being a person means. It is very different depending on whether we can have freedom in the libertarian sense or not. The loss of libertarian freedom would be destructive to personhood in the ways it was discussed in chapters 3-5. If the arguments presented there are correct, then this is a major destruction that would rob us of most things we value in being persons.

The issue of rationality and intellectual responsibility

To the causal theory the most damaging of the conclusions of the early chapters concerning the values which are available or not, depending on whether we can be free in the libertarian sense or not, was the one concerning rationality. Philosophers may live with the idea that they are self-determining creatures only in a shallow sense, and that they are not really morally blameable or appraisable for anything they do or achieve. But it is hard to imagine a philosopher who can live with the idea that he is not in the position to advocate any thesis in an intellectually responsible way. If my arguments in chapter 5 are sound then this alone makes the philosophical price of giving up on libertarian freedom unbearable. The choice, then, is between accepting the extreme non-reductive ontology of persons or giving up on responsible philosophical discourse completely

Unfortunately, my arguments in the fifth chapter rested on two assumptions, the categorical difference between the norms of rationality and causal laws, and the nomological nature of causation, and for the latter I did not argue. I plan to make the argument complete later by supplementing an argument for the nomologicality of causation.

It should be noted, though, that even if causation can be anomalous, rationality can be predicted of a thought produced by a thinking mechanism only if interactive property dualism is assumed, contradicting the physicalist dogma of the causal completeness of physics. There is no reason why a causal theorist of freedom (a compatibilist) could not be an interactive property dualist, but if the causal theory (compatibilism) is viewed as part of a broader

naturalistic philosophical agenda, i.e. naturalizing freedom, then this result in itself could be uncomfortable for most of its proponents.

Appendix: Some Major Interpretations of Quantum Mechanics, Determinism and Absolute Simultaneity

The basic elements of the formalism

In quantum mechanics the state of a physical system is represented by a wave-function ($\Psi(\mathbf{r},t)$)[303] or a state-vector $|\Psi\rangle$[304]. The two representations are equivalent, as the set of all possible wave-functions, with the operations of addition and multiplication with real numbers, form a linear vectorspace, more precisely, a Hilbert space of denumerably infinite dimensions, complete, separable and equipped with a scalar product, of which every wave-function is a vector.

Now it is worth to pause for a minute to appreciate the significance of this very fundamental feature of the formalism. The fact that the set of possible states is represented mathematically with a vectorspace means that if Ψ_1 and Ψ_2 are two possible states of a system, then any linear combination of these two states, $c_1\Psi_1+c_2\Psi_2$, is also a possible state of the system. To be vivid, if a system can be in state Ψ_1, and it can be in state Ψ_2, then it can also be in a state which is half-Ψ_1 and half-Ψ_2 ($(1/\sqrt{2})\Psi_1+(1/\sqrt{2})\Psi_2$). Now if we think that macroscopic phenomena are realized by an underlying quantum mechanical reality, then this feature of the formalism will lead into a serious conceptual difficulty. For the supposition that if an object can be in state 1 and in state 2 then it can also be in a state which is the combination of states 1 and 2 is contrary to our experience (think of a cat that can be alive and dead, but not half-alive and half-dead). This conflict between the additivity of states and macroscopic experience, taken together with the linearity of the dynamical law (to be introduced in a minute) that governs the deterministic evolution of quantum mechanical states, gives rise to the nonepistemic (i.e. objective) character of the probabilistic nature of quantum mechanical description of reality, unless the formalism is

---

303 This is a twice continuously differentiable and square-integrable function of time and spatial co-ordinates, which maps them onto the set of complex numbers.

304 The more abstract vector formalism was developed by Werner Heisenberg. It is often called the 'Heisenberg picture', in contrast to the wave-function representation, which is called the 'Schrödinger picture'.

supplemented with an addition that makes the emergence of definiteness of classicality out of the indefiniteness of quantum mechanics deterministic, if such an addition is possible. And this linearity is also what is responsible for the holism and non-locality that emerges from the correlatedness of the states if spacelike separated, non-interacting constituents of composite systems, such as those involved in EPR-type scenarios, which feature is exploited by some theorists of time arguing for absolute simultaneity, despite the contrary predicament of the special theory of relativity.

All of this we will discuss shortly. For now, let us continue with the basics of the formalism.

Measurable physical properties are represented mathematically as Hermitian operators operating on the Hilbert space of possible state-vectors (($\hat{\boldsymbol{O}}(\Psi)$ $\boldsymbol{P}(\Psi)$, $\boldsymbol{Q}(\Psi)$, etc., representing physical properties *O, P, Q*, etc.).[305]

For every physical property there are physical states such that when a system is in one of them, the measurement of the physical property in question has a definite single possible outcome. These states are called the eigenstates *of that particular property*.[306] When a system is in an eigenstate of a particular property, then it is mathematically represented with the following equation

$$\hat{\boldsymbol{O}}\Psi = \boldsymbol{\lambda}\Psi.$$

This equation expresses the fact that the operator representing the property being measured, if the state of the system is an eigenstate of that property, assigns a wave-function to the wave-function representing the state of the system, which differs from the latter only inasmuch as it is multiplied with a certain number, or, it may be said perhaps more instructively in the vector representation, that the operator representing the given property sends the vector representing the state of the system to a real-number multiple of

---

[305] Hermitian operators are a special (self-adjoint) class of linear operators, which are essentially functions that assign vectors to vectors. Self-adjoint operators can be rendered as linear combinations of projectors, i.e. operators that project vectors orthogonally into subspaces of the Hilbert space, leaving vectors that are already in that subspace unaffected. The formalism could be presented by using projectors only. The use of projectors would be particularly useful when discussing some modern approaches to quantum mechanics, such as the consistent histories approach of Griffiths, Omnès, Gell-Mann and Hartle, but on this we will touch upon only very briefly.

[306] The representations of these states are called eigenfunctions or eigenvectors.

itself.[307] This number, λ, is called an eigenvalue, and it represents the measured value of the given property.

One of the phenomena that called for a new theory was that many physical properties appeared to have a discrete set of possible values when measured. Now this phenomena is represented in the formalism of quantum mechanics by the fact that Hermitian operators can have a discrete set of eigenfunctions (eigenvectors), with a discrete set of eigenvalues (usually called the "spectrum") belonging to them, and with the assumption that whenever a physical property is measured, the outcome of the measurement is one of the eigenvalues, even if the system was not in an eigenstate at the start of the measurement process.

At the time of the measurement the system can indeed be in a state which is not an eigenstate. However, since the eigenvectors of Hermitian operators form a basis in the Hilbert space of possible states,[308] whatever state a system might be in, it can be obtained as a linear combination of the eigenstates of the property in question:

$\Psi = \sum c_i \Psi_i$, where $\hat{O}\Psi_i = \lambda_i \Psi_i$, for any Ψ and $\hat{O}$.

Since these non-eigenstates can be obtained as linear combinations of eigenstates, i.e. as the sum of eigenstates added together with different weights, they are also called "superposition states".

When a system is in a superposition state with respect to the property that is being measured, the formalism gives only a probabilistic prediction of the outcome of the measurement, provided that the state of the system is known. The probability that the i-th possible outcome (λ_i) will come out as the result of the measurement depends on the weight with which the eigenstate corresponding to this possible outcome figures in the linear combination that gives out the state in which the system actually is, when it is being measured:

$P(\lambda_i) = |c_i|^2$.

This equation is called the Born rule.

Between any two measurements physical systems are thought to evolve continuously and deterministically, in accordance with a

---

[307] The operator assigns a vector to the original vector which is aligned exactly the same way in the vectorspace, only its length is different.

[308] Just like three orthogonal vectors form a basis in normal three-dimensional space.

dynamical law, which, in the non-relativistic case, is the time-dependent Schrödinger equation (in the relativistic case it is the Dirac or the Klein-Gordon equation). Upon measurement the smooth evolution of the state-function is broken, and it instantaneously (and, if there is nothing more to be said about the causal evolution of the system, indeterministically) collapses into one of the eigenstates.[309] In his groundbreaking work about the foundations of quantum mechanics John von Neumann called these two phases of the evolution of the quantum state "process 2" and "process 1", respectively.[310] The result of the measurement will be the eigenvalue corresponding to the eigenstate into which the superposition state that is the result of process 2, up to the time of the measurement, collapses in process 1.

There is a lot more to it, but this is how it essentially works.

The Measurement Problem and the Interpretations of Quantum Mechanics

The problem arises from the duality of the dynamical processes described above, von Neumann's processes 1 and 2. The puzzle about them is how nature should know when to switch from process 2 to process 1. Of equal right we may also say that the problem arises from the fact that possible quantum states form a vectorspace. That is why process 1 had to be introduced to make sense of the fact that we never encounter superposition states observationally, although they are, in principle, possible.

Process 2 is the smooth evolution of the wave-function governed by the dynamical law. Process 1 is the collapse of the wave-function in measurement. But what is measurement, and what is so special about it? From nature's point of view the measuring device is just a physical system that gets entangled with the system to be measured. Why isn't it that the state of the whole entangled complex just evolves in accordance with the dynamical law? Why does it collapse, instead?

---

[309] So the idea is that when a definite measurement result is obtained, the measured system becomes to be in an eigenstate of the measured property. This principle is called 'the eigenvalue-eigenstate link'. To every eigenstate corresponds a subspace of the Hilbert space. When this subspace is one-dimensional, the eigenvalue is called "nondegenerate", when it has more than one dimensions, it is called "degenerate". The obtaining of the eigenvalue may be represented also as the state being projected down to the subspace of the Hilbert space corresponding to the eigenvalue.

[310] Von Neumann 1955.

But if it did not collapse, the puzzlement would perhaps be even greater.[311] Suppose that the system to be measured is in an eigenstate of the operator representing the observable to be measured in the measurement process. Let this eigenstate be denoted by Ψ_i , where $\hat{\boldsymbol{O}}\Psi_i = \lambda_i\Psi_i$, $\hat{\boldsymbol{O}}$ being the operator representing the physical property to be measured, λ_i being its i-th possible value.

At the start of the measurement process, the measuring apparatus is in an initial state to be denoted by χ_0. The initial state of the whole complex, before the system to be measured and the measuring device start interacting with each other, is $\Psi_i \otimes \chi_0$.[312]

Given the linearity of the dynamical law, there is a linear operator that transforms the initial state of the whole complex into its final state, if the smooth evolution governed by the dynamical law is all that happens between the beginning and the end of the measurement process:

$$\boldsymbol{L}(\Psi_i \otimes \chi_0) = \boldsymbol{\Phi}_i.$$

$\boldsymbol{\Phi}_i$ is the final state of the whole entangled system when the measurement device reads λ_i.

Now what if the initial state of the system to be measured is not an eigenstate, but a superposition state, which is a linear combination of eigenstates: $\Psi = \sum c_i\Psi_i$, and there is no collapse involved in the measurement process?

The initial state of the whole entangled complex is then

$$(\sum c_i\Psi_i) \otimes \chi_0,$$

and its final state is

$$\boldsymbol{\Phi} = \boldsymbol{L}((\sum c_i\Psi_i) \otimes \chi_0).$$

Given the linearity of $\boldsymbol{L}$,

---

311 I am indebted for this formal presentation of the problem to László E. Szabó.

312 Tensor product, represented by the symbol ⊗, can be though of to the analogy of a simple logical conjunction. Ψ_i and χ_0 are functions of different variables. Before the measurement process starts the values of these different variables are supposed to have no effect on each other. The symbol ⊗ means that we simply unite the Hilbert-spaces in which they dwell, and look at a wave-function that we obtain by the logical conjunction of Ψ_i and χ_0, now inhabiting all the dimensions Ψ_i and χ_0 inhabited before separately.

$$\Phi = \sum c_i \mathbf{L}(\Psi_i \otimes \chi_0) = \sum c_i \Phi_i.$$

Well, it is the state of the whole entangled complex when the measuring device reads... what? Surely, when we look at it, we never see the measuring device to be in a superposition state. If it is a reliable measuring device, appropriate for the purpose for which we are using it, we never see it reading more than one value, or none at all, as the result of the measurement process. Rather, we see it to read one definite value of the measured physical observable (its pointer always points at one direction at one time, it doesn't split), which, if the eigenvalue-eigenstate link is retained, signifies that the entangled complex is in one of the eigenstates at the end of the measurement. Thus, presumably, there must be something more to the measurement process than just the smooth evolution guided by the dynamical law.

Schrödinger's cat makes this problem vivid. Or deadly. Suppose that in a closed container there is a cat, a cyanide capsule, an electron source, a Stern-Gerlach device, and a mechanism linked to it that breaks or not breaks the capsule, depending on the result of the spin-measurement. Had the wave-function of the electron, which is initially a superposition of the spin-up and the spin-down eigenstates, not collapse, the cat would have to be in a superposition of cat-alive and cat-dead states. And it seems absurd.

But when and why does the wave-function collapse? Is it when someone looks into the container to see if the cat is still alive or not? In that case it would be thought to be a case of collapse-upon-measurement, in which the cat serves as the display of the measuring device. But maybe it is absurd to think that a cat could be in a half-dead, half-alive state even if no one ever looked into the container. If so, then even if one does have a look at the cat, it is not this that triggers the collapse of the wave-function, it must have taken place earlier, because of the cat. But a cat is a fishy object, because it might have some sort of consciousness, so maybe it counts as an observer. Or maybe the collapse doesn't have to do with the cat either. Maybe it has to do with the fact that the braking of the cyanide capsule is a "macroscopic" event, meaning that there is always a definite truth about whether it happened or not.[313] But what does it take to be

[313] Actually, von Neumann argued that the formalism that involves both process 1 and process 2 gives the same predictions for observations irrespective of where we place the

"macroscopic"? Is it simply being big? How big? Why should there be a physics for small things, and another for big things? Could it be that only small things have a wave-function, and that it collapses whenever a small thing interacts with a big thing? It is not easy to believe. A single proton is a quantum object. So is a hydrogen atom. It is a quantum mechanical two-body problem. A molecule involving the hydrogen atom is probably a quantum mechanical multi-body problem. So is a system of two interacting molecules. And so is a system of three. And four. And... What happens when a system of interacting molecules gets "big"—acquires a size of, say, a centimetre, i.e. if a quantum-mechanical multi-body problem realizes a macroscopic object? Does it seize to be a quantum-mechanical multi-body problem? Or does "macroscopic" mean only that the thing can be observed, so it is evidently absurd to suppose that it is in a superposition state, because we see it isn't, regardless of its size (or mass, or the number of atoms it contains)? But then we are coming close again to the metaphysically surprising supposition that it is the presence of an observing mind, after all, that collapses the wave-function, and we have to suppose that superposition states are possible when no one is looking.

This is the measurement problem. Another name for it could be "the macro-objectification problem".[314] To sum up, the problem is that either we have to embrace the collapse of the wave-function, but then we have to justify it somehow, or, alternatively, we have to explain away the apparent contradiction between the definiteness of observable phenomena and the "superposition-ness" of the wave-function, which it retains if it is thought to evolve always according to the linear dynamical law, without collapsing.

Both broad ways of solving the measurement problem have several variations. These are called the "interpretations" of quantum mechanics.[315]

"cut", i.e. when we think process 2 gives way to process 1, so models involving different hypotheses about when the collapse takes place are empirically equivalent. This feature of the theory is sometimes referred to as "the moveability of the von Neumann cut".

[314] One way of putting what is so puzzling about quantum mechanics would be to say that it predicts with extreme accuracy and reliability what we are to observe, but leaves us in complete puzzlement about what it is that we observe, i.e. in what sort of fundamental ontology it is realized and how the observed reality emerges from this fundamental ontology. It is in this context that the measurement problem can be called the problem of macro-objectification. (See the introduction of Ghirardi 2007.)

[315] Some of what are usually called "interpretations" of quantum mechanics should be more appropriately called "remakes" of quantum mechanics, since they not only

Is it consciousness that collapses the wave function? – Von Neumann and Wigner

Historically, perhaps the most radical way of solving the measurement problem came close to being standard first, i.e. making sense of the collapse of the wave-function by effectively denying the mind-independent evolution of objective reality.

The supposition that the mind has a part in the creation of reality was actually made by John von Neumann already.[316] Drawing on his work, other founding fathers of quantum mechanics, most famously another Hungarian, Eugene Wigner, went as far as to claim that the laws of nature cannot be formulated without reference to the mind, provided that the laws of nature are essentially quantum mechanical:

> When the province of physical theory was extended to encompass microscopic phenomena, through the creation of quantum mechanics, the concept of consciousness came to the fore again: it was not possible to formulate the laws of quantum mechanics without reference to the consciousness. All that quantum mechanics purports to describe are probability connections between subsequent impressions (also called 'apperceptions') of consciousness, and even though the dividing line between the observer, whose consciousness is being affected, and the observed physical object can be shifted towards one or the other to a considerable degree, it cannot be eliminated.[317]

---

interpret but also modify the formalism to some extent. (For example, the Everett interpretation is a clear case of interpretation, whereas the GRW interpretation is clear case of remake. Both will be discussed below.) Remakes in principle allow for empirical tests, since the alterations in the formalism (for example the addition of any non-linear term to the dynamical equation of the standard theory – as in the case of the GRW interpretation – may result in altered predictions in some cases, although all remakes were originally designed to reproduce the predictions of the standard theory to the largest extent possible. The crucial tests to distinguish between the remakes and the standard theory are out of our present technological reach, but may become achievable in the future. Keeping this in mind, in the sequel I will refer to both interpretations and remakes as interpretations.

316 Ibid. Chapter VI.

317 "Remarks on the Mind-Body Problem", 1961, reprinted in Wigner 1967, pp. 171-84. Werner Heisenberg held a similar opinion (1958). One would perhaps assume that this extravagantly mentalistic view was characteristic only of the early days of quantum mechanics and has already died out, but it hasn't. An example for a recent prominent advocate would be Stapp (1993, 2007).

This is one way of making sense of the collapse: whenever a subsystem of reality is forced to squeeze out of itself an answer to a question posed by a macroscopic observer that is unambiguously interpretable for him, its quantum state collapses. (A question can be posed by simply looking at something.) Besides positing an irreducible mind-body dualism (of which Wigner speaks very explicitly), which many philosophers (unlike me) find unacceptable in the first place, this interpretation of quantum mechanics faces other serious difficulties, too. For one, as far as our best biological and cosmological theories go, this planet, and presumably the whole known universe, lacked any conscious observer for quite a while. It is hard to believe that the first collapse happened only after the first conscious being evolved. And even if the universe had to wait for the first quantum collapse to take place until the first conscious being appeared in it, how conscious this being had to be? It is unclear that there is a sharp enough criterion for being conscious in the required sense. Does a cat qualify, for example?[318] Another problem is that, although there is no evidence to the contrary, it is really hard to believe that the pointer of a measuring device points at a definite value on its scale only when someone looks at it (or, perhaps, that a measuring device itself is conscious), or, to put it more generally, it is hard to believe that there is no problem with macroscopic superposition states as long as no one is looking.

The dissatisfaction with holding the observer (and perhaps also the measuring apparatus) metatheoretical can be summarized as David Bohm did:

> If the quantum theory is to be able to provide a complete description of everything that can happen in the world...it should also be able to describe the process of observation itself in terms of the wave functions of the observing apparatus and those of the system under observation. Furthermore, in principle, it ought to be able to describe the human investigator as he looks at the observing apparatus and learns what the results of the experiment are, this time in terms of the wave functions of the various

[318] As John Bell wrote: "Was the wave function waiting to jump for thousands of millions of years until a single-celled living creature appeared? Or did it have to wait a little longer for some highly qualified measurer – with a PhD?" (1981, p. 611).

> atoms that make up the investigator, as well as those of the observing apparatus and the system under observation. In other words, the quantum theory could not be regarded as a complete logical system unless it contained within it a prescription in principle for how these problems were to be dealt with.[319]

Collapse theories without metatheoretical observers – the unified dynamics of Ghirardi, Rimini, and Weber

But there is a variety of ways of justifying of the collapse without positing a metatheoretical status to the observer.

The most worked out, most widely discussed, and maybe the most widely accepted, spontaneous collapse theory is the GRW (or GRWP) account, which is an elaborate version of the view that the collapse into eigenstate in cases like that of Schrödinger's cat has to do with the system's being big, after all.[320] Ghirardi, Rimini, and Weber (and Pearle) modified the dynamical law by adding stochastic and non-linear terms to the original equation of the standard theory, to make it sensitive to the number of particles involved, or, in a subsequent version, to the average particle number within an appropriate volume, thus, at bottom, positing a mechanism whose effect is negligible for microscopic systems, but highly relevant for macroscopic ones. The result is a unified dynamical theory, which accounts for microscopic systems the same way as the standard theory does; for micro-macro interactions such as measurements without the difficulties which arise if we assume the interaction of the measurement apparatus and the measured probe to be governed by a linear dynamical equation; and for the classical behaviour of macroscopic objects. In micro-macro interactions this unified mechanics leads to the non-linear and stochastic collapse of the wave-function. The mechanism which is responsible for the reduction of the quantum state of a system grows in effectivity as we move from micro to macro systems. As John Bell noted, this theory

---

319 1952, p. 583.

320 Ghirardi, Rimini, Weber, "Unified Dynamics for Microscopic and Macroscopic Systems", 1986. The best introduction to this theory for non-specialists I am aware of is Giancarlo Ghirardi's entry on the topic in *The Stanford Encyclopedia of Philosophy* (Ghirardi 2007).

> allows electrons (in general microsystems) to enjoy the cloudiness of waves, while allowing tables and chairs, and ourselves, and black marks on photographs, to be rather definitely in one place rather that another, and to be described in classical terms.[321]

And this without reference to the observing mind.

What these solutions to the measurement problem (von Neumann-Wigner, and GRW) have in common, is that they are realist about the wave-function, hold that the description of the state of a physical system with the wave-function is complete, and they all embrace an objectively indeterministic instantaneous collapse of the wave-function of possibly extended systems, which invokes absolute simultaneity.

There are interpretations of quantum mechanics, however, which choose the other broad way of dealing with the measurement problem. They believe in the unitary (and deterministic) evolution of the state of physical systems, unbroken by collapses, and attempt to account for why it appears so as if collapses were happening, i.e. why it is that we see Schrödinger's cat to be either alive or dead, although its superposed state does not collapse.

Bohm's non-local hidden variable theory

One of the types of theories of this breed started its career very early on, with the work of Louis de Broglie.[322] The idea, which was probably based on a hint that originally came from Einstein, was advocated for a short period by both de Broglie and Max Born, but then has been abandoned as they converted to the Copenhagen interpretation. The abandoned idea has been taken up by David Bohm[323], resulting in a theory which gives a strikingly simple solution to the macro-objectification problem. The simplicity of the solution is best demonstrated with how the theory deals with the particle-wave duality involved in phenomena like the classical two-slit experiment.

---

321 Bell 1986, p. 364. Cited by Ghirardi 2007.
322 De Broglie 1928.
323 Bohm 1951. About the history of the de Broglie-Bohm theory see Goldstein 2007.

In his very lucid introduction to Bohm's theory[324], Sheldon Goldstein cites two remarks Richard Feynman made about the significance of the two-slit experiment. Feynman wrote that

> [the two-slit experiment is] a phenomenon which is impossible, absolutely impossible, to explain in any classical way, and which has in it the heart of quantum mechanics. In reality it contains the only mystery.[325]

At another place Feynman adds that

> [the experiment] has been designed to contain all of the mystery of quantum mechanics, to put you up against the paradoxes and mysteries and peculiarities of nature one hundred per cent. ... How does it really work? What machinery is actually producing this thing? Nobody knows any machinery. Nobody can give you a deeper explanation of this phenomenon than I have given, that is, a description of it.[326]

Well, de Broglie and Bohm thought there was more to give than just a description.

The phenomenon to be interpreted in the two-slit experiment is that when an electron leaves a mark on a screen that it reaches by going through either, or both, of two slits of a wall separating the source from the screen, it behaves as a small particle having a definite position, whereas, when a sufficiently large number of electrons are sent through the slits, the marks they leave on the screen build up an interference pattern that is characteristic of a wave coming through both slits at the same time, the two parts of it interfering with each other in the space between the slits and the screen, *even if only one electron is emitted towards the slits and the screen at one time*, and a subsequent electron is emitted only when the previous one has reached the screen. The mystery is effectively the difficulty of being realist about an entity that is both a spatially extended wave and a pointlike particle at the same time.

---

[324] Goldstein 2007.
[325] Feynman, Leighton, Sands 1963, p 37.
[326] Feynman 1967, p. 130 and p. 145.

De Boglie's simple solution is that there are in fact two entities, associated with each other. An electron is a combination of a wave and a particle, and both are real. The particle is guided by the wave. The wave is the quantum mechanical wave-function that evolves according to the Schrödinger equation, the particle is a classical entity, traversing along a definite deterministic trajectory, having both a definite position and a definite momentum at all times. The trajectory of the particle is determined by its initial position and by a "quantum potential", analogous to potentials used in classical dynamics, derived from the phase of the wave-function, which is a complex periodical function of the spatiotemporal co-ordinates. Regarding its choice between the two slits, and its subsequent trajectory from the slit to the screen, the movement of each electron is determined by its initial position and the quantum potential determining its velocity. The quantum potential, in turn, is determined by the wave, which does go through both slits, and does interfere with itself subsequently.

This is a hidden variable theory whose hidden variable is the initial position of the electron. It is hidden because it cannot be known. But if it is assumed that its probability-distribution is $|\psi|^2$, where ψ is the initial wave-function of the electron, then the de Broglie-Bohm trajectories of an ensemble of electrons, the distribution of whose initial positions are thought to conform with this assumption[327], give out the interference pattern seen in the experiment.

Quite contrary to Feynman, John Bell commented on this explanation of the two-slit experiment that the only mystery about it is why it was so generally ignored, given that it is so natural and simple.[328]

Bohm generalized the idea and the formalism to systems of many particles, in which case we have an ontology consisting of the particles and a wave, which is the wave-function of the quantum mechanical many-particle problem. The wave evolves according to the Schrödinger equation that accounts for the interactions within the system, and it guides all the particles. Quite pictorially, an N-particle system can be thought of as a definite point in a 3N-dimensional configuration space that is being pushed around by the flow of the probability distribution derived from the wave-function of the system the standard way, just as a massless particle would be in a compressible fluid. The flow of the fluid is determined by the

---

[327] Known as the "quantum equilibrium hypothesis".
[328] Bell 1987, p. 191.

Schrödinger equation, the effect of the flow on the point representing the configuration of the N-particle system is determined by the guiding equation.[329] The theory proved empirically correct in all non-relativistic experimental situations, which is little surprise, as it was deliberately designed to reproduce the predictions of standard quantum mechanics.

It is an advantage, however, over the standard theory that the "macro-objectification problem" simply does not arise, given that, according to Bohm's theory, all particles have a definite position at all times. There are no two distinct laws for the evolution of the wave-function, one for measurement situations, and one for normal situations. The wave-function evolves always deterministically according to the Schrödinger equation. It never collapses. So there is no need to explain why measurement situations would be special. They aren't special.

Yet, as the theory is designed to reproduce the predictions of standard quantum mechanics, it has to account for how it is that in post-measurement situations systems continue their evolution as if their wave-function had collapsed into an eigenstate. The answer Bohm gave to this question is effectively an early version of the now widely endorsed theory of decoherence. Decoherence, in general, is an account of how components (or the interference between components) of a superposed wave-function become irrelevant to the description of the system due to the system's interaction with the environment, without a collapse taking place. In Bohmian mechanics the component of the superposed wave-function that corresponds to the actual outcome of the measurement becomes the only one that is relevant for the post-measurement evolution of the system, because it is the only component which significantly differs from zero at the well-defined locus of the system in the configuration space, so it is the only component that effectively guides it. In the rather large configuration space including all degrees of freedom of the measured system, the measuring apparatus, and all systems in the environment with which they interact, other components of the post-measurement wave-function have little chance to overlap with the one that corresponds to the measured state. (Intuitively, the larger is the number of dimensions of a vectorspace, the larger is the probability that two randomly picked vectors will be orthogonal.) So for all

[329] Barrett 2003.

practical purposes the complex post-measurement wave-function can be replaced with the eigenstate corresponding to the outcome of the measurement. This feature of Bohmian mechanics is called the "effective collapse" of the wave-function in measurement situations.

It should be noted that Bohmian mechanics not just eliminates the irreducible reference to an observer or a physical situation that is specifically a measurement, as do also spontaneous collapse theories, but also introduces parameters, i.e. the actual configuration of the system, with which the statistical and indeterministic account of quantum mechanical phenomena by the standard theory is made deterministic and complete. It is not done, however, exactly the way it was envisaged by Einstein, Podolsky and Rosen. They thought it was the non-locality involved in EPR-type phenomena, which could not be explained away unless the existence of hidden variables was assumed, that signified that the standard description must have been incomplete, and that there must have been more to quantum reality than the information contained in the wave-function. Bohm's theory is a hidden variable theory, yet, the addition of the hidden variables does not do away with the non-locality of the standard theory. (In fact, it was Bohm's theory that led Bell to the proof that no local hidden variable completion of quantum mechanics was possible.) The non-locality of Bohmian mechanics is very explicit. Since the holism of entangled many-particle wave functions is inherited from the standard theory to Bohm's, and since the velocities of the particles are determined, through the guiding equation, by the wave function, the velocity of any one member of a many-particle system will manifestly depend on the positions of all the others, however far they might be, as long as they are entangled, and the effect of a change in the position of a distant particle on the velocity of a particle here will be instantaneous. As a result, although it does not invoke instantaneous collapses of wave-functions of spatially extended systems, since the interdependence between its hidden variables, i.e. determinate particle positions, is instantaneous action at a distance, Bohmian mechanics is no less dependent on a frame-independent notion of simultaneity as are collapse theories.

The multiple versions of the Everett multiverse

There is another family of interpretations, however, which solves the measurement problem, like Bohmian mechanics, by getting rid of

the collapse of the wave function, but which is free from the explicit instantaneous action at a distance dynamics of Bohmian mechanics, too, so hold out hope to eliminate the threat quantum mechanics poses to the special relativistic symmetry of inertial observers. Or so its inventor, Hugh Everett boldly claimed:

> Fictious paradoxes like that of Einstein, Podolsky and Rosen which are concerned with such correlated, non-interacting systems are easily investigated and clarified in the present scheme.[330]

Now, if Everettian quantum mechanics realizes this hope, then the purchase of the argument from quantum mechanics to absolute simultaneity is inversely dependent on the reasons we might have to accept the Everett interpretation.

As it is the case with many theories that are exceptionally attractive in some respects, there is a price to pay for the nice features. In the case of the Everett interpretation the price to pay for a solution to the measurement problem which is possibly local is a really extravagant metaphysics, which in itself is enough for many to be reluctant about considering Everettian approaches seriously. To complicate the issue a little further, the question whether the Everettian interpretation is really local may depend on the exact kind of extravagance that is involved, and it varies from one version to another.

On all variants of the Everett interpretation, which Everett himself called the "relative-state formulation" of quantum mechanics, or the "theory of the universal wave-function", the wave-function is not merely a means to encode information about reality, but it is thought to be real, independently of any observer, in fact, it is thought to model the *only* fundamental reality. It evolves always smoothly, as prescribed by the same dynamical law applicable in all physical situations. This assumption amounts to dropping von Neumann's process 1 from the orthodox formulation of quantum mechanics, leaving Everett with the task of explaining how it is that we get determinate results in measurements, despite the fact that the measured systems, more often then not, start interacting with the measuring apparatus and the observer in superposed states, and that

---

[330]Everett 1957, Section 5. (The article is accessible online at http://www.univer.omsk.su/omsk/Sci/Everett/paper1957.html.)

the dynamical law that governs the evolution of the whole entangled system is linear. The answer to this question in Everett's own words is this:

> [W]ith each succeeding observation (or interaction), the observer state "branches" into a number of different states. Each branch represents a different outcome of the measurement and the *corresponding* eigenstate for the object-system state. All branches exist simultaneously in the superposition after any given sequence of observations. The "trajectory" of the memory configuration of an observer performing a sequence of measurements is thus not a linear sequence of memory configurations, but a branching tree, with all possible outcomes existing simultaneously in a final superposition with various coefficients in the mathematical model.[331]

The numerous different Everettian interpretations of quantum mechanics are in fact attempts at making sense of this passage. The idea is quite simple though. The superposed wave-function does not collapse into any of the eigenstates. At the end of the measurement process, the wave-function of the complex consisting of the measured system, the measuring apparatus, and the observer, is the superposition of the states into which the state of the complex system would have evolved, had the measured system been in each of the eigenstates of which its actual pre-measurement superposition state was composed. The linear evolution of the wave-function of the complex system during the measurement process has, however, an important effect. Whereas in the pre-measurement state the components of the superposition state interfered with each other, by the time the measurement process is over, there is no interference any longer between the components. The measured system-measuring apparatus-observer complex is not in any of the eigenstates after the measurement, so the observer is not in the state of recording one particular eigenvalue corresponding to the measured system's being in one particular eigenstate, but he is in the state of recording *each* particular eigenvalue *relative to* the system being in the corresponding eigenstate. The obtained results are all real. The different relative

---

[331] Ibid.

post-measurement states of the measured system-measuring apparatus-observer complex are in fact different components of the universal wave-function, which do not interfere with each other and thus are reidentifiable over time, and function as different branches of reality which are mutually inaccessible from each other.

Now it seems that what we are facing here is an interpretation of quantum mechanics that solves the measurement problem at the cost of requiring the Lewisian reality of all possible worlds[332] that have the same physics (the same dynamical equation) and grew out of the same initial conditions (the same initial universal wave-function), or at least the reality of the Parfitian fission of observers.[333]

The exact metaphysics of the branching varies from one version of the Everett interpretation to the other. It seems though that something like the Parfitian splitting obtains in all of them. An important dividing line between different families of Everettian interpretations is whether they want mental states (beliefs about observed measurement outcomes) to supervene fully on physical brain states. If yes, then something like the Lewisian plurality of worlds is inevitable besides the Parfitian weirdness of personal identity through time. This is how it is in the words of Lev Vaidman:

> "I" am an object, such as Earth, cat, etc, "I" is defined at a particular time by a complete (classical) description of the state of my body and of my brain. "I" and "Lev" do not name the same things (even though my name is Lev). At the present moment there are many different "Lev"s in different worlds (not more than one in each world), but it is meaningless to say that now there is another "I". I have a particular, well defined past: I correspond to a particular "Lev" in 2002, but I do not have a well defined future: I correspond to them all. Every time I perform a quantum experiment (with several possible results) it only seems to me that I obtain a single definite result. Indeed, Lev who obtains this particular result thinks this way. However, this Lev cannot be identified as the only Lev after the experiment. Lev before the experiment corresponds to all "Lev"s obtaining all possible results. Although this approach to the concept of personal identity seems

[332] Lewis, D. 1986.
[333] Parfit 1986.

> somewhat unusual, it is plausible in the light of the critique of personal identity by Parfit 1986. Parfit considers some artificial situations in which a person splits into several copies, and argues that there is no good answer to the question: Which copy is me? He concludes that personal identity is not what matters when I divide.[334]

Now there is no need for artificial situations. Splitting happens on a regular basis every time when collapse theorists would record a collapse.

As it figures in Vaidman's paragraph, to every "I" there is a corresponding "world". It is because he wants the mental state corresponding to the "I" to supervene on a physical state, also corresponding to the "I", which physical state is part of a larger physical context. In every quantum experiment not only the sentient beings, which may be involved in it as observers, but the whole world of material objects and properties split into equally existing but slightly different copies that occupy the same space and the same time, or which, together, occupy a branching spacetime. These branching worlds correspond to components of the universal wave-function which do not interfere with each other any longer. The universal wave-function corresponds to the Universe encompassing the totality of these parallel worlds. Versions of the Everett interpretation in which the worlds multiply, as well as observing consciousnesses do, are called Many Worlds Interpretations (MWI). MWI's come in several different variations, the classic ones are most prominently that of DeWitt[335] and Graham[336], in which worlds actually split, as it was stated here, or that of David Deutsch[337], in which quantum experiments do not exactly split worlds, but rather distinguish between subsets of a pre-existent infinite ensemble of worlds, which he calls the 'Multiverse'.[338]

There are Everettian views, however, on which the different mental states of sentient beings corresponding to their recording of

---

[334] Vaidman 2002.
[335] De Witt 1971.
[336] Graham 1973.
[337] Deutsch 1996a.
[338] The main motivation for this is to give a meaning to the probabilistic predictions of quantum mechanics, as we will see a little later. It is also a convenient feature of this version, in comparison to DeWitt's, that there is no worry about the violation of conservation laws when worlds split.

different macroscopically determinate measurement outcomes do not fully supervene on different physical states. David Albert and Barry Loewer[339] accept, contrary to collapse theories, that the wave-function always evolves as prescribed by the linear dynamical law, and, contrary to Bohm, that the universal wave-function gives a complete description of physical reality, but, contrary to MWI, do not embrace the idea of splitting, or multiple, physical worlds. In Vaidman's dual talk of himself as "Lev" to which several "I"'s correspond, only "Lev" figures at the physical level, physical reality is single. The different "I"'s are mental and are associated to the same physical state. As far as only physics is concerned, this interpretation is very faithful to the unitary evolution of the wave-function, and faithful in a strictly minimalist way: no collapse, no hidden variables, no branching or splitting, only unitary evolution, and this is all that there is to say about physics, including the physics of the brain. Since it adds nothing to the quantum mechanical skeleton at the physical level, this theory was called "the bare theory" by Albert.[340] In consequence of the bareness of their theory, however, Albert and Loewer have to reconcile two claims that, on psychophysical reductionist grounds, do not fit together, i.e. that the physical state of an observer is typically a superposition state, and that the mental state of the same observer, having mostly well-defined beliefs (e.g. about outcomes of measurements), typically isn't. What the reconciliation requires is a manifest mind-brain dualism and going Everettian about minds. In the Albert-Loewer Many Minds Interpretation (MMI) every physical observer state is assumed to be linked to a continuous infinity of irreducibly mental (non-physical) minds. And whereas the physical state of the brain is thought to evolve continuously and deterministically as governed by the dynamical law, it is assumed that the evolution of minds is discontinuous and genuinely probabilistic. Definite macroscopic belief states are obtained in measurements *as if* they would correspond to actual eigenstates into which the wave-function would collapse with a probability determined by the Born rule if the orthodox theory was true. For every measurement the Born rule determines the chance for each mind, and so the proportion of the infinite collection of minds, to obtain a certain determinate result, and to evolve indeterministically to the mental state corresponding to

339 Albert and Loewer 1988.
340 Albert 1992.

it. This is how the phenomenology and the empirically firm predictions of standard quantum mechanics are retained.[341]

Now MMI comes in several other versions, too. There are ones that are even more economical in respect of the number of mind-independent physical worlds than Albert and Loewer's. Albert and Loewer have one. A radical, largely Berkeleyan, version of MMI, which has been developed recently by Cambridge physicist Matthew J. Donald, has actually none.[342] Donald proposes that individual minds and their structures should be the fundamental entities of our ontology. These minds are experiencing and processing definite (classical, macroscopic) information. The material world is merely appearance to these minds. The minds are obeying strict laws, which determine their possible experiences and their probabilities. These laws *are* the laws of physics for Donald. Nothing happens to these

---

[341] On the Albert-Loewer theory the evolution of each mind is *probabilistically guided* by the unitary physical evolution of the wave-function. One might come up with the idea that there is an alternative to this interpretation with just one mind per brain, whose evolution is deterministic. This alternative could be construed to the analogy of the Bohmian pilot wave theory whose 'beable' being piloted by the wave would be the definite belief state of the mind, rather than the actual definite configuration of the system. (There is an important difficulty though: the determinate belief state of the observer is not necessarily about the definite configuration about the system, whereas in Bohmian mechanics configuration is fundamental. If, however, there is a good explanation from the Bohmian's part of how he can account for the definiteness of all kinds of experience with his deterministic theory using a particular preferred basis, i.e. configuration, then possibly that argument can be applied, mutatis mutandis, to this case, as well.) It is hard to see though any advantage that this theory could claim over Bohm's. (Certainly, they would be non-local in exactly the same way.)

But even without thinking of a possible analogy with Bohm's theory, the question arises whether the supposition of a multitude of minds is really inevitable in Albert's and Loewer's theory. Why can't the unitary evolution of the physical state (the bare theory) be supplemented with the indeterministic evolution of a single mind, instead of an infinity of minds, if the superposition of the mental state on the physical state is broken anyway. (This proposal was considered by Albert 1992.) Of course, this would not be a version of the Everett interpretation any longer. But Albert's and Loewer's many minds theory is already a departure from Everett's original intention, whose main motivation was that he couldn't believe that the theory should respect a fundamental distinction between observers and non-observers. Everett wanted to account for definite macroscopic states as existing objectively, independently of minds being aware of them. Although the reference to many worlds is not to be found in any of his published works, David Deutsch (1996b) reports that in conversations Everett defended his theory in terms of parallel universes. But setting aside the issue of being faithful to Everett, the single-mind variant faces the problems of its own, I mean the "mindless hulk" problem (Albert 1992, p. 130) and its consequences, that we will discuss a few pages below.

[342] Donald 1990, 1995, 1997, 1999. See also his website at: http://www.bss.phy.cam.ac.uk/~mjd1014/index.html.

minds that is not prescribed by these laws, nor has the consciousness of these minds any influence on the course of events. So in this sense Donald, being ontologically an idealist monist, can be said to be a physicalist. The physics governing the evolution of minds is the unitary linear evolution of quantum states. The problem of the coherence between the usually superposed character of the quantum mechanical state, representing now possible experiences, on the one hand, and the definiteness of actual experiences, on the other, is solved the Everettian way, by introducing many minds experiencing each of the components of superposition states separately. The problem of how a macroscopically definite mind-independent reality would arise from an underlying mind-independent microscopic reality, which is quantum mechanical and so smeared, does not arise, since there is no mind-independent reality. (Neither has Donald to face the problem of having false beliefs about physical reality, like Albert and Loewer do, as we will see shortly, for the same reason.)

Speaking of different MMI's, it should be noted that Michael Lockwood's much discussed many minds theory is not exactly a version of MMI, as the latter was defined here. Lockwood does believe in the multiverse (much in the same way as Deutsch). He believes that the multiplicity of physical reality at large is "an inescapable consequence of allowing superpositions of what classical physics would regard as mutually exclusive alternatives"[343], of which the multiplicity of minds, which are physical systems, is a special case. Deutsch explains Lockwood's reason to prefer to present his theory under the label of "many minds", rather than "many worlds" or "many universes". It is

> because of the classical connotations of the word 'universe'. He points out that the picture of the multiverse as being simply a collection of entities each of which is similar to the universe of classical physics, misrepresents some important features of the multiverse's structure. In particular, describing the multiverse in terms of different, incompatible sets of observables slices it into different, inequivalent sets of 'universes'. So...the 'layering' structure of the multiverse as a whole is highly arbitrary. By contrast, the 'layering' structure for states of mind (given that they

[343] Lockwood 1996. Cited in Deutsch 1996b.

> are associated with certain observables) is in principle unique.[344]

The picture proposed here is that even if there is no objective, non-perspectival answer to the question of how the multiverse should be layered, it is a physical fact that there is more to the multiverse than what there is in any of the layers, and that it extends to infinitely many parallel ones.

Now this passage from Deutsch reveals why one might be motivated to opt for many minds instead of many worlds—apart, of course, from the consideration that the former might perhaps be thought to do somewhat better when facing Occam's razor. It is because of the preferred basis problem.

The preferred basis problem arises because the Hilbert space formalism of quantum mechanics allows for the decomposition of the wave-function into the linear combination of eigenstates in many (in fact, infinitely many) ways. Now this problem is really pressing, because decomposition into eigenstates, on many worlds theories, is really a decomposition into worlds. So there must be an answer to the question why this basis of the Hilbert space, rather than that. Reasons to prefer a basis to others may arise from the fact that what we are ultimately explaining is the definiteness of experience, and as Jeffrey Barrett put it,

> [m]aking the total angular momentum of all the sheep in Austria determinate by choosing such a preferred basis to tell us when worlds split, would presumably do little to account for the determinate memory I have concerning what I just typed.[345]

But those who are philosophically inclined to prefer MWI to MMI, for they want mental states to supervene on physical states, are perhaps after a theory that does not rely on the observer, and on what is relevant for his perceptual capacities, in any irreducible way. If one is to place the mind within the confines of the world of physical systems, then there is no reason to favour the splitting of worlds over the collapse of the wave-function, if it is ultimately the consciousness that is responsible for both.

---

[344] Deutsch 1996b.

[345] Barrett 2003.

On many minds theories, however, the problem of the preferred basis is legitimately solved by appealing to the interest of the mind. As Deutsch said above, "the 'layering' structure for states of mind (given that they are associated with certain observables) is in principle unique." Many minds have an advantage over many worlds if the experiences of sentient beings can be definite (and those beings might have evolved to survive utilizing their definite experience) without the unambiguous decomposition of the universal state to definite macrorealms at the physical level. Of course the many minds theorist is then required to propose a theory as to the emergence of definite experience without a physically definite macrorealm, but if the Everettian approach is applied only at this level, then the perspective or the interest of perceiving minds is perhaps legitimately invoked, and so the problem of the preferred basis is no longer so pressing. Now if within the context of many worlds, instead of many minds, the actual "*layering structure of the multiverse*" can be only perspectival, then this problem is equally well escaped. The question is how it can be only perspectival, if the *multiplicity of physical reality* is "an inescapable consequence of allowing superpositions".

Another important problem that Everettian interpretations may encounter is the problem of probability. Frankly, the problem is that the theory is deterministic, so it has no *alternative* futures. Take the example that the world splits into two branches. The fact is that both branches will be real, and I will be in both of them. What would it mean to say that branch A is realized with 99% probability, and branch B with 1% probability, or, for that matter, that with 99% probability I will find myself in branch A, and with 1% probability I will find myself in branch B?

> "The problem amounts to this: it seems that the concept of probability can only apply to a situation given that only one...out of a range of possibilities...is true, or is realized, or actually occurs, so as to exclude all the others; precisely what Everett denies."[346]

Now this is a really important problem, because quantum mechanics does predict probabilities, and if we cannot make sense of those probabilities, then we strip quantum mechanics from its testable

---

[346] This is how Simon Saunders summarizes the worry on p. 2 of his 1998.

predictions that made it empirically so robustly confirmed. In Maudlin's words:

> [S]ince there are no frequencies in the theory there is nothing for the numerical predictions of quantum theory to mean. This fact is often disguised by the choice of fortuitous examples. A typical Schrödinger-cat apparatus is designed to yield 50% probability for each of two results, so the 'splitting' of the universe in two seems to correspond to the probabilities. But the device could equally be designed to yield a 99% probability of one result and 1% probability for the other. Again the world 'splits' in two; wherein lies the difference between this case and the last?[347]

If there is a genuinely random process, over and above the unitary deterministic evolution of the universal wave-function, such as the random evolution of minds on Albert and Loewer's MMI, then of course the interpretation of probability may be unproblematic. If, however, one prefers to stick to the supervenience of the mental states of observers on their physical states, then this option is not available.

The most beautiful theory proposed to deal with both of these problems I know of, largely along the lines of Deutsch's "multiverse" with "a perspectival layering structure", is the relational theory that Simon Saunders put forward in a series of articles in *Synthese* on "Time and Quantum Mechanics", from 1995 to 1998.[348]

The core idea of Saunders's theory is that the "collapse" of the wave-function and the notion of "actuality", in relation to the multiplicity of quantum mechanical possibilities, are comparable to the "passage" of time and the notion of "the present". Just like the

---

[347] Maudlin 1994, p. 5. In a different but equally instructive formulation of the problem from Loewer the relation of the problem of probability to that of identity over time is made explicit: "Prior to measuring the x-spin of a z-spin electron, a rational observer... ought not to have a degree of belief ½ that she will observe spin up. Either she will think that this degree of belief is 0 because she will not exist at the later time or, if she identifies herself with all the minds associated with her brain at the end of the measurement, she will believe that at the conclusion of the experiment she will certainly perceive that x-spin is up and also she will believe that x-spin is down and so assign a degree of belief of 1 to each of these." (1996, p. 230.)

[348] Saunders 1995, 1996, 1998. Perhaps the most easily accessible overview of Saunders's theory is given in his 2000 article. David Wallace proposed a very similar theory (2002).

passage of time is not a metaphysical reality, and temporal determinations obtain only as "B-determinations", i.e. only as relations of events, in the special theory of relativity, the quantum state does not really collapse, and the value-definiteness of observables obtains only as a relational property on the interpretation of quantum mechanics that Saunders is advocating, which, he claims, "is not just compatible with relativity" (for it will turn out to be local), "but...shares its spirit, too".[349] What he suggests is essentially that the Minkowski spacetime and the Hilbert space of physical states, in this respect, should be treated on a par.

The Everettian relative-state formulation of quantum mechanics is presented as an analogue of the McTaggartian B-theory of time:

> The connection should be quite plain: the difficulties of attribution of tense [or "A-determinations", in McTaggart's terminology], given a space-time description, mirror the difficulties of attributing definite values to observables, given a superposition of eigenstates. It seems that either every event is "present", or that none is, just as every eigenvalue or none is "actual". The method of solution is the same in each case: whilst "event E is now", and "event E' is now" are contradictory, given E and E' occur at different times, introducing new events W, W' we obtain: "event E is now relative to event W", "event E' is now relative to event W"", and there is no longer a difficulty. Likewise "observable X has value x", and "observable X has value x"" are inconsistent for x and x' distinct, but introducing a new observable Y we may say instead: "observable X has value x relative to value y of Y", and "observable X has value x' relative to value y' of Y", and there is no longer a contradiction.[350]

The analogy clarifies the status of what Deutsch called "the multiverse", and its "layering structure".

> [T]he parallel with tense is particularly helpful; for one, it makes clear that what is involved in the Everett procedure is poorly made out in terms of a set-theoretical collective of

[349] Saunders 1995 p. 4.
[350] Ibid, p. 2.

> worlds ("Everett-worlds"). Space-time may be understood as an infinite collection of 3-dimensional worlds, but not in the sense that the total mass or energy is additive.[351]

Or as Jeremy Butterfield wrote interpreting Saunders's view:

> [J]ust as someone who accepts the tenseless conception of time can readily accept instants i.e. spacelike slices of spacetime as (i) useful or even indispensable for describing phenomena, and yet (ii) not any substantive ontological commitment additional to the commitment to spacetime, so also an Everettian can readily accept worlds as (i) useful or even indispensable, and yet (ii) not a substantive commitment additional to the commitment to actuality's being described by the universal state.[352]

Whereas on the many minds approaches the preferred basis is explicitly defined in terms of a set of minds performing specific observations, and whereas on the more traditional versions of the many worlds approach a distinguished set of subsystems need to be specified which work as apparatuses to obtain definite values for specific properties, and the rest of the world is in a relative state for these subsystems being in eigenstates, in the picture Saunders is proposing the basis of the Hilbert space, relative to which the multiverse decomposes into value-definite worlds (or "layers"), is just similar to a frame of reference, relative to which Minkowski spacetime decomposes into space and time.

Given that decomposition into macrorealms is a relative matter, relative ultimately to the interest of sentient beings (in this respect this view resembles many minds theories), and as such it has to be definite only "for all practical purposes", that is, a small amount of vagueness is not excluded on conceptual grounds, the definition of the basis relative to which value-definiteness is to be obtained can, as Saunders puts it, be "the business of decoherence theory", which Saunders frames in the formalism of the consistent histories approach of

---

[351] Ibid. The mentioning of the additivity of mass or energy refers to an objection to DeWitt's actually splitting worlds variant of many worlds theory, i.e. that it violates conservation rules, which is dominantly considered fatal.

[352] Butterfield 2002, p. 34. (The page number refers to the preprint available in the Los Alamos Archive.)

Griffiths, Omnès, and Gell-Mann and Hartle[353]. On this theory the "suppression of interference" between disjoint histories (time-ordered products of Heisenberg-picture projections which project the universal state vector onto subspaces of the Hilbert space that correspond to definite values of certain observables) in the consistent history space (which in this picture play the role of the collection of worlds or branches or layers of the multiverse) is a wholly natural physical process the account of which itself makes no mention of "observers" or "measurements", and which is governed by the unitary dynamics. The preferred subsystems (substituting the "apparatuses" of the older theory) are all the systems that decohere interacting with the environment.

(The concept of decoherence was already touched upon passim in relation to Bohmian mechanics. It is a process quite similar to dissipative processes in thermodynamics, in which a system which has a small number of (relatively massive) degrees of freedom interacts with an environment of a very large number of (light) degrees of freedom. A paradigm example would be a dust particle interacting with air molecules and photons. The interference between different eigenstates as of the few degrees of freedom of the system (e.g. the co-ordinates of the position of the centre of mass of the dust particle) is suppressed by the interaction with the environment (so the state of the composite system, that is the quantum state of the system consisting of the dust particle + its environment, develops into a superposition of non-interfering components in which the dust particle has a very nearly definite position whose evolution is very nearly classical).[354])

The analogy between tense and value-definiteness is extensively used also in Saunders's solution to the problem of probability. As he explains, the problem can be solved by "extending the relational account of tense to modal attributes".[355] In his view, the problem of probability as an objection against the Everett approach is much like as though special relativity would be criticized for the lack of change in it.

---

353 Cf. Griffiths 1984, Omnès 1994, Gell-Mann and Hartle 1995. Gell-Mann and Hartle relate decoherence to the emergence of stable and definite experience which makes it possible for "INGUS"es (information gathering and utilizing systems) to survive in their environment.

354 Cf. Zurek 1991, or Zeh 1970.

355 Saunders 1998, p. 3.

> Consider...the more or less instinctive criticisms that have been directed to each [i.e. STR and Everett]: if space-time as a whole is unchanging, then it cannot describe change (if the universal state as a whole is deterministic, then it cannot describe probability); if there is no such thing as "time-flow", then the distinction between past, present and future is unreal (if there is no such thing as state-reduction, then value-definiteness is illusory). [...]
>
> If tenseless relations are adequate to the treatment of tense, then so too are deterministic relations adequate to indeterminism. Temporal facts do not come to be true in time; probabilistic facts are not made true by chance.
>
> In both cases there is the underlying conviction that something "essential" has been omitted; certainly the intuitive notions of "time-flow" and "actualization" no longer enter at the level of foundations. But it need not be claimed that these notions are empty or meaningless: it may be that part of their meaning can be recovered at other levels in the development of the theory, remote from the first principles.[356]

This approach makes the proliferation of worlds characteristic of some Everettian views even relative to others, most prominently of those of Deutsch and Lockwood, unnecessary. On these views each component of the universal wave-function (state-vector), after it has been decomposed in a preferred basis, is inhabited not by one world but rather an infinite population of identical worlds, of which more is suggested to continue to possible alternative future A than to B if and only if A is more probable than B. This proposed excessive ontology is discarded by Saunders as an outmoded attempt to ground quantum mechanical probability in the principle of the a priori equiprobability of a pool of outcomes, which, he claims, never really worked even in more conventional cases.[357] Instead, Saunders defines probability as a relational property of pairs of events, not dependent in any way on objective becoming, or objective indeterminacy, but grounded measure-theoretically using the Hilbert space norm as the measure,

---

[356] Saunders 1995, p. 3.

[357] See Section 6 of his 1998.

allowing him to have only one consistent history (world) per component.[358] He says:

> Indeterminacy lies in the future. In physics probability habitually involves time. Typically, the concept of probability applies to states of affairs qua future, in relation to the present. Correspondingly, probabilities are conditional, they are *de facto* relations.[359]

And:

> Here one has a space of all possible histories, where each history is considered as a 4-dimensional whole, without any preferred foliation. A measure is defined on this space, and the histories are required to satisfy the consistency condition with respect to it (essentially that distinct histories do not interfere with one another). We suppose that one of these histories is ours, and that it is "typical", as defined by this measure. Conditional probabilities for A, conditional on B, can then be defined as the measure of all those histories which contain B and A, divided by the measure of all those which contain B.[360]

Let us come back now to the question of locality and relativistic covariance. As it was declared by Everett, it was within the purview of the Everett interpretation to make quantum mechanics local and relativistically covariant. Now there is a lively debate about whether this goal has been achieved by any of the variants that we have outlined.

The various Everettian approaches *prima facie* seem to suffer from the same difficulty right at the outset. It is that the relative state was defined as instantaneous, in terms of relations that hold between simultaneous states of components of spatially extended entangled systems, or, as we might equally say, it was defined as a component of the universal state at an instant.

But as it stands, it only means that the relative state is defined relative to a foliation of spacetime, which should not be a problem as

---

[358] See Section 3, ibid.
[359] Ibid., p. 9.
[360] Saunders 2000, pp. 5-6.

long as this foliation is not unique, that is, as long as any other spacelike foliation would also be allowed to be chosen for this purpose, and the relative states defined using different foliations are related by the appropriate symmetry transformations. This problem in itself is not insurmountable.[361]

To be more specific about the issue of non-locality one has to go into the specific features of the various versions of the Everett interpretation for the fulfilment of the Everettian promise of a local quantum theory depends on the metaphysical setup in which the Everettian idea of coexisting observed realities is actually cashed out. It is enlightening to start with a theory, Albert and Loewer's, in which the supervenience of mental states on physical states is imperfect. Meir Hemmo and Itamar Pitowsky[362] argued that Albert and Loewer's theory is local (relativistically covariant), and that this feature of the theory (i.e. nonsupervenience) is essential for its locality. Guido Bacciagaluppi, however, in an article prompted by Hemmo and Pitowsky's,[363] argued that among the theories that have the perfect supervenience of the mental state on the physical state, Simon Saunders's is perfectly local, contrary to what Hemmo and Pitowsky say. In what follows I will briefly review these two partly conflicting arguments, having regard also to what Albert and Loewer and Saunders themselves said on the matter.

The "bareness" of the Albert-Loewer theory is a great advantage in respect of locality. On this theory the different relative states do not correspond to coexisting physical worlds, physical reality remains single in measurement-like situations, so we do not have to account for how the branching or splitting of worlds should be thought of in a locality-respecting way. The lack of any extra structure additional to the unitary evolution of the universal state, allows the theory to be perfectly local and covariant as far as physics goes, as the dynamical law can be written in a relativistically covariant form.

Given the dualism of the account, i.e. the imperfect supervenience of the mental state on the physical state, non-locality, of course, can reappear at the mental level. In order to avoid the recurrence of non-locality at the level of minds, the preferred basis that determines the

---

361 Cf. Butterfield 2002.

362 Hemmo and Pitowsky 2003.

363 Bacciagaluppi 2002. (Although Bacciagaluppi's paper was first to appear in print, the Hemmo-Pitowsky paper had been available well before that on-line. Bacciagaluppi remarks that his paper was prompted by Hemmo and Pitowsy's.)

expansion of the universal state in terms of brain eigenstates and their relative states is thought of as local, i.e. as a set of local bases, one for each observer, rather than a global preferred basis for all brains in the universe. This makes it possible for the evolution of minds belonging to different observers to be independent from each other, guided probabilistically only by the reduced state of the brain of the given observer.

This alone, however, would not yield locality at the mental level. The feature of the theory that a multitude of minds is associated with each observer is also vital.

It is easy to see that leaving unchanged the Albert-Loewer assumption concerning the genuinely probabilistic evolution of minds, and the independency of the evolution of minds of different observers, but positing only one mind per observer (instead of a continuous infinity of them) Albert and Loewer would run into a curious difficulty concerning the experiences of observers in the two wings of the EPR experiment, who are later allowed to meet and communicate their experiences to each other. This difficulty is called the "mindless hulk" problem.[364] The problem arises because if there is only one mind per brain, then only one component of the post-measurement superposition state of the brain is actually tracked by the indeterministic evolution of the mind, leaving the other component mindless, and given the independence of the evolution of the minds of different observers, in the EPR experiment the trajectories taken by the two minds would fail to anticorrelate with probability .5, so there would be probability one half that an observer will later encounter a brain in his branch of the superposition state without an inhabiting mind.

If Albert and Loewer would seek to repair this situation by letting the trajectories of the two minds correlate, i.e. by giving up the principle that the evolutions of minds are independent from each other, and as such are guided only by the reduced state of the brain to which they belong, then, as Bacciagaluppi points out, they would have to embrace the idea that the evolution of correlated minds is guided by the reduced state of the composite system consisting of all the brains involved, and that the preferred basis defining the options for the indeterministic evolution of minds is holistic. This, however,

---

[364] Albert 1992, p. 130, Bacciagaluppi 2002, Section 4, Hemmo and Pitowsky 2003, Section 2.

would amount to the reintroduction of the non-locality, of which the bare theory got rid at the physical level, at the mental level.

The only alternative is to make sure that all possible mental trajectories are actually taken by minds. That is why Albert and Loewer have the infinitely many minds. If every mind belonging to the two brains in the two wings of the experiment takes either of the two trajectories with probability .5, then an infinite pool of such minds secures it that neither of the trajectories will be mindless.

If a mind later meets a brain that performed the experiment in the other wing he will find it inhabited with a right kind of mind, that is, when he meets the observer from the other wing, the other observer will report that he is in the memory state of remembering that he recorded the opposite result in the spin measurement. For this, no non-local co-ordination of individual mind-trajectories is required, because this fulfilment of the expectation concerning the report of the observer from the other wing is guaranteed by the evolution of the physical state, given (i) that the post-measurement physical state is a superposition of states with matching physical brain eigenstates, (ii) that by the Everettian assumption all post-measurement subsystem-states are allowed to interact only with relative subsystem-states (subsystem-states in their own branch), and (iii) that the evolution of the physical state is the unitary evolution, guided by a relativistically covariant law. Both branches of the superposition are inhabited, and the probability that the report of the left-wing observer will match with the quantum mechanical expectation of the right-wing observer, or vice versa, is 1, even though there is no dependence between the evolutions of individual minds.[365]

The Albert-Loewer interpretation is therefore covariant at both the physical and the mental level.[366]

---

[365] Albert 1992, p. 132, Hemmo and Pitowsky 2003, Section 4, Bacciagaluppi 2002, end of Section 4. How to apply then the Bell theorem to this situation? Albert and Loewer say that there are no matters of fact about the correlation of the outcomes of the measurements in the two wings (that is, the correlation of the belief-states of minds in the two wings, just having performed the measurement), so there is nothing to which the Bell theorem could be applied. Correlations obtain only between the expectations of one observer and the reports of the other, and such correlations are local.

[366] Hemmo and Pitowsky argue that there is a weaker sense of non-locality even in this case, arising from the correlation between *subsets of minds* in the two wings. If, as Albert and Loewer say, there is no fact of the matter about the correlatedness of individual mind-trajectories, there is no fact of the matter about whether or not a report of having measured an up-spin, say, in the left wing of the EPR experiment by a left-wing observer, upon meeting with a right-wing mind who measured a down-spin, corresponds

Hemmo and Pitowsky argue that, unless one gives up on the reidentifiability of minds over time, the genuinely indeterministic evolution of minds, characteristic of Albert and Loewer's theory, is also essential for the theory to be local in Bell's sense. The indeterminism of the evolution of minds requires the supervenience of the mental state on the physical state to be imperfect. If one insists on the full supervenience of the mental on the physical, one has to think of the evolution of minds as deterministic. But in that case the evolution of minds cannot be independent from each other in EPR-like situations, as they are determined by spacelike separated correlated physical events.

But if this is so, if the locality of the Everett picture can be bought only at the price of giving up on the supervenience of the belief-state on the brain state and the relative state of the rest of the world, including the pointer of the measuring device and the measured system in a measurement situation, then the Everett interpretation saves locality not merely by adopting a dualistic metaphysics but also by adopting a really grim view on empirical knowledge.

Even if one is willing to accept the dualism of the Albert-Loewer view, the latter one may find very discouraging. The nonsupervenience involved in Albert and Loewer's theory is not the kind of nonsupervenience dualists usually want to secure for the mental realm. This nonsupervenience means that the link between the physical state that obtains at the end of the measurement process and the belief state of the observer about it is broken. The observer's belief that he obtained a determinate measurement result will always be false.

This is quite embarrassing in itself, but this is not all. It seems that if this is true, then we may run an Epicurean argument on the bare theory, largely analogous to the one Epicurus offered against determinism, the one we discussed in detail in chapter 5. Consider,

---

to a mind who has actually observed an up-spin in the left wing. In short, from the perspective of the down-spin right-wing mind, there is no fact of the matter whether or not a left-wing up-spin *report* (which is a physical thing) corresponds to an up-spin left-wing mind. Hemmo and Pitowsky say this is analogous to the mindless hulk problem, so that if that was a problem, so is this. The problem can be avoided only by assuming that there is a correlation between the sets of minds of the two observers. But, as they argue, this non-local correlatedness of subsets of minds is a feature of the uncollapsed quantum state, so the evolution of the minds need not depend on a preferred frame of reference, so this weaker non-locality does not affect relativistic covariance.

say, Albert's belief state that he believes that the bare theory is true. Now Albert may have some sort of justification for this belief, but it is quite certain that he doesn't have an *empirical* justification for it. Well, what about a scientific theory that tells of scientific theories (itself included) that they cannot be justified empirically? Isn't it a shot in the foot?[367]

However, Bacciagaluppi argues that it was too quick, and that locality can be saved also if one sticks to supervenience. He says his preferred Everettian approach, that is, Simon Saunders's, is perfectly local.

If one insists on the supervenience of the belief-state on the physical state of the brain, which is correlated with the relative state of the rest of the world, then one has as many physical realities as components in the superposition state. Now, in respect of locality, two questions arise about these realities.

The first concerns whether these realities are defined in a relativistically meaningful way. If yes, then, these realities cannot be thought of as three-dimensional totalities, containing every object (relative to a belief-state of an observer) at a given time, but they are four-dimensional totalities, containing the whole histories of such totalities of objects, for only the latter is relativistically invariant.

If the 'worlds' of a many worlds theory are construed as consistent histories, as it is the case with Saunders's theory, then this criterion is met. As it is emphasized by Bacciagaluppi, it is an important feature of Saunders's theory that the consistent histories are thought of as temporally ordered sequences of projections which are local in the sense of axiomatic algebraic quantum field theory.[368] It means that the projections that correspond to the events in the consistent histories are members of local subalgebras of operators associated with open bounded regions of Minkowski spacetime. The axioms of algebraic QFT include the requirement of the commutativity of operators supported by spacelike separated regions, which practically means that locality is enforced by the axioms of the theory. As a result

> The 'events' in each history will be embeddable in a Minkowski space-time. Identifying histories with Everett

[367] Cf. Barrett 2003.

[368] Cf. Haag 1996. For a brief overview see Section 2 of Earman and Ruetsche 2005.

> worlds, Everett worlds will thus have a Minkowski space-time structure.[369]

So far so good.

The other question is whether the splitting of worlds growing out of a measurement-like situation is thought of in a relativistically meaningful way.

Quite naturally, if the splitting of worlds is thought of to the analogy of the immediate collapse of the wave-function, i.e. as happening globally, affecting a whole totality of objects inhabiting a simultaneity plane, then the theory cannot be relativistically covariant.

Now, in Saunders's theory, measurement-like situations, not necessarily actual measurements involving an apparatus and an observer, but situations in which the interference between different components of the global state is suppressed, obtain when decoherence interactions take place.

So decoherence interactions need to be local, and then the splitting of worlds is local, too. So the axiom of the commutativity of operators corresponding to decoherence interactions that take place in spacelike separated regions should be thought of as a restriction on the range of decoherence interactions that can actually happen.

As a consequence, the effective stochastic process of the transition into either of the decoherent branches growing out of a decoherence interaction should be though of like this:

> The process is to be defined at each space-time point, as an effective state reduction on a certain 3-dimensional surface in space-time. But this surface is not a time-slice, a space-like hypersurface. It is the surface of the forward light-cone of each point. As such, on making a Lorentz-transformation, this surface, and the associated data is *invariant*. … Just as important, these stochastic processes at these space-time points are all *independent* of each other.[370]

So the spacetime occupied by splitting Everettian worlds is a branching spacetime, but the branching is not to be thought of as taking effect at an instant in respect of everything that exists at that instant, which would mean that the branching takes place along a

---

[369] Bacciagaluppi 2002, p. 117.

[370] Saunders 2000, p. 11. Emphasis in the original.

simultaneity plane, but, rather, spacetime branches always along the future lightcone of a decoherence event.

Again, it is uncontroversial that splitting or branching so conceived is a relativistically covariant notion.

Now what about the EPR experiment?

What Saunders says about the issue is essentially that EPR-type correlations simply do not arise. The only correlations to account for are between the expectation of the observer in one wing of the EPR experiment concerning the report of the observer in the other wing, and the actual report of the observer in the other wing. And that is local.

Indeed, in the framework put forward by Saunders EPR-type correlations *cannot* arise:

> [W]e could never in this way obtain EPR-type correlations: the independence of the local processes implies factorizability, and that in turn the Bell inequalities (which we now to be violated). But that is only to say once again: the Everett relational approach is fundamentally local; it is consistent with relativity, as goes the observed phenomenology, only in this light. These correlations between measured outcomes are incorporated into *local* records; and the amplitudes for such records, in which correlations do not obtain, are vanishingly small.[371]

What shall we make of this?

It is clear that in the EPR experiment there are two decoherence events: the measurements that take place in the two wings. So spacetime branches along the two respective future lightcones.

But what happens to spacetime at the intersection of those two lightcones?

If experimenters in the two wings would have been measuring two *different* uncorrelated components of the spin of the two electrons, then the intersection of the two lightcones would be a divergence surface for a further split: both of the two worlds diverging along the two lightcones would split again, so, on the whole, we would have four worlds, inhabiting four leaves of spacetime.

---

[371] Ibid.

It would mean that the worldline of the observer in the left wing of the experiment, who measures, say, the x-spin of the left electron, branches at the point representing the event of the measurement. It splits into two, corresponding to the possible outcomes of the measurement. From this point on let us follow the worldline of the self of the left observer who measured the x-spin to be up. When this worldline reaches the future lightcone of the measurement event in the right wing it branches again. Again it splits into two, corresponding to the two possible reports an observer from the right wing, whom the observer from the left wing now can encounter, can give about whether he measured, say, the z-spin of the right electron to be up or down. The same goes for the worldline of the observer from the right wing. Both worldlines will split into four in two steps. Quite unsurprisingly, this account we have given of the branching caused by two noncorrelating measurements referred only to notions, events, lightcones and worldlines, which are frame-independent, that is, Lorentz invariant.

Now in the case when the two observers in the two wings of the experiment measure the same spin-component of the two electrons, say, the z-spin, the only difference is that there is no second branching along the intersection of the future lightcones of the two measurement events. The worldlines of the two observers, which split at the point representing the measurement they perform, do not split again when they reach the future lightcone of the measurement in the other wing. The spacetime which the decoherent histories involving the two measurement events inhabit will have only two leaves.

What Aspect and his collaborators verified in 1981, strictly speaking, is this lack of further branching along the intersection of the two lightcones. And, it seems, it can be accounted for in a Lorentz covariant language.

Here is the great advantage of the Everettian approach over collapse theories, that is, why an Everettian split can be local while a collapse cannot, as Bacciagaluppi explains it:

> In the context of a collapse theory, collapse along the future light cone is not admissible, because in the case of perfect correlations the collapses on the future light cones of the measurement events would not necessarily match up. If they did, the collapse would be along the piecewise lightlike surface connecting the two spacelike separated

> events. In the present context, however, (effective) collapse along the future light cone can account for the phenomena, because all leaves of the space-time are inhabited.[372]

This completely local account sufficiently explains perfect correlations (anticorrelations) observed within any particular spacetime leaf:

> [I]f there is no further splitting and the space-time has just the two leaves associated with the components |+>|-> and |->|+>, the mere fact that all leaves of the space-time are inhabited by some branch of an observer (if any leaf is) suffices to explain why different observers have matching results if they meet in any one leaf.[373]

So we have a version of the Everettian approach in which observation states fully supervene on physical states, yet the account is fully local, so it provides no ground for an argument for absolute simultaneity.

---

[372] Bacciagaluppi 2002, p. 119.
[373] Ibid.

References

Adams, R. M. 1985. "Involuntary Sins", *Philosophical Review* 94, pp. 3-31.

Albert, D. 1992. *Quantum Mechanics and Experience*. Cambridge MA: Harvard University Press.

Albert, D., Loewer, B. 1988. "Interpreting the Many Worlds Interpretation", *Synthese* 77, pp. 195-213.

Anscombe, G. E. M. 1981. "A Reply to Mr C. S. Lewis's Argument that 'Naturalism' is Self-Refuting", in *The Collected Papers of G. E. M. Anscombe II: Metaphysics and the Philosophy of Mind*. Oxford: Blackwell.

Aspect, A., Ph. Grangier, G. Roger 1981. "Experimental Test of Local Realistic Theories via Bell's Theorem", *Physical Review Letters*, 47(7), pp. 460-3.

Armstrong, D. M. 1973. *Belief, Truth and Knowledge*. Cambridge: Cambridge University Press.

-------- 1961. *Perception and the Physical World*. London: Routledge and Kegan Paul.

Augustine 1986. *The Confessions of Saint Augustine*. Brewster: Paraclete Press.

Ayer, A. J. 1954. "Freedom and Necessity", in Ayer, *Philosophical Essays*. New York: St. Martin's Press. Reprinted in, and cited from Watson 1982.

Bacciagaluppi, G. 2002. "Remarks on Space-time and Locality in Everett's Interpretation", in J. Butterfield, T. Placek (eds.) *Non-locality and Modality*. (NATO Science Series II. Mathematics, Physics and Chemistry, Vol. 64.) Dordrecht: Kluwer Academic Publishers, pp. 105-22.

Baker, L. R. 1988. "Cognitive Suicide", in H. Grimm, D. D. Merill (eds.) 1988.

Balfour, A. J. 1989. *Theism and Humanism*. Germantown: Periodicals Service Co.

Barrett, J. 2003. "Everett's Relative-State Formulation of Quantum Mechanics", in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy (Spring 2003 Edition)*, URL = <http://plato.stanford.edu/archives/spr2003/entries/qm-everett/>.

Bell, J. S. 1987. "How to Teach Special Relativity", in *Speakable and Unspeakable in Quantum Mechanics*. Cambridge: Cambridge University Press, pp. 67-80.
-------- 1986. "Six Possible Worlds of Quantum Mechanics", in *Proceedings of the Nobel Symposium 65: Possible Worlds in Arts and Sciences*. New York: de Gruyter.
-------- 1981. "Quantum Mechanics for Cosmologists", in C. J. Isham, R. Penrose, D. W. Sciama (eds.), *Quantum Gravity 2: A Second Oxford Symposium*. Oxford: Clarendon Press.
-------- 1966. "On the Problem of Hidden Variables in Quantum Mechanics", *Reviews of Modern Physics*, 38, pp. 447-52.
-------- 1964. "On the Einstein Podolsky Rosen Paradox", *Physics*, 1, pp. 195-200.
Bernecker, S., F. Dretske (eds.) 2000. *Knowledge: Readings in Contemporary Epistemology*. Oxford: Oxford University Press.
Boghossian, P. A. 1990. "The Status of Content", *The Philosophical Review*, Vol. 99, No. 2, pp. 157-184.
-------- 1989. "The Rule-following Considerations", *Mind*, 98.
Bohm, D. 1957. *Causality and Chance*. London: Routledge and Kegan Paul.
Bohm, D., Aharonov Y. 1957. "Discussion of Experimental Proof for the Paradox of Einstein, Rosen, and Podolsky", *Physical Review Letters*, 108, 1070.
Bok, H. 1998. *Freedom and Responsibility*. Princeton: Princeton University Press.
Broad, C. D. 1952. "Determinism, Indeterminism and Libertarianism", in C. D. Broad, *Ethics and the History of Philosophy*. London: Routledge & Kegan Paul.
Butterfield, J. 2002. "Some Worlds of Quantum Theory", in R. Russell, J. Polkinghorne et al. (ed.), *Quantum Mechanics (Scientific Perspectives on Divine Action vol 5)*, Vatican Observatory Publications, pp. 111-140. Preprint available in both the Los Alamos archive: quant-ph/0105052, and the Pittsburgh University Philosophy of Science Archive: 00000204.
Callender, C. forthcoming. "On Finding 'Real' Time in Quantum Mechanics", in W. L. Craig, Q. Smith (eds.). *Absolute Simultaneity*. New York: Oxford University Press.
Campbell, C. 1951. *In Defense of Free Will*. London: Allen & Unwin.
Carnap, R. 1963. "Intellectual Autobiography", in Schlipp, P. A., ed. *The Philosophy of Rudolf Carnap*. LaSalle, Illinois: Open Court.

Cassirer, E. 1920. *Zur Einstein'schen Relativitätstheorie*, Berlin: Bruno Cassirer Verlag.
Chisholm, R. 1966. "Freedom and Action", in Lehrer, ed. (1966).
Churchland, P. M. 1984. *Matter and Consciousness.* Cambridge, Mass: MIT Press.
-------- 1981. "Eliminative Materialism and the Propositional Attitudes", *Journal of Philosophy* vol. 78, no. 2.
Churchland, P. S. 1987. "Epistemology in the Age of Neuroscience", *Journal of Philosophy* 84, pp. 544-52.
-------- 1986. *Neurophilosophy: Toward a Unified Science of the Mind/Brain.* Cambridge, Mass.: The MIT Press.
Craig, W. L. 2001. *Time and the Metaphysics of Relativity*. Dordrecht: Kluwer Academic Publishers.
Crane, T. 2002. *Elements of Mind.* Oxford: Oxford University Press.
Davidson, D. 1980. *Essays on Actions and Events.* Oxford: Clarendon Press.
-------- 1974. "Psychology as Philosophy", reprinted in Davidson 1980.
-------- 1963. "Actions, Reasons, and Causes", reprinted in Davidson 1980.
De Broglie, L. 1928. "La nouvelle dynamique des quants", in H. Lorentz (ed.), *Electrons et Photons: Rapports et Discussions du Cinquieme Conseil de Physique Solvay.* Paris: Gauthiers-Villars.
Dennett, D. C. 2003. *Freedom Evolves.* London: Allen Lane The Penguin Press.
-------- 1995. "On Giving Libertarians What They Say They Want", in O'Connor 1995.
-------- 1991. *Consciousness Explained.* Boston: Little, Brown and Company.
-------- 1990. "The Myth of Original Intentionality", in Mohyeldin Said, K. A. et al. (eds.) (1990), pp. 43-62.
-------- 1988. "Evolution, Error and Intentionality", in Wilks, Y., Partridge, D. (eds.) 1988.
-------- 1987. *The Intentional Stance,* Cambridge, MA: MIT Press/A Bradford Book.
-------- 1984a. *Elbow Room.* Cambridge, MA: MIT Press.
-------- 1984b. "I Could Not Have Done Otherwise—So What?", *The Journal of Philosophy* LXXXI, 10, pp. 553-67.
-------- 1978. *Brainstorms.* New York: Bradford Books.

Deutsch, D. 1996a. *The Fabric of Reality*. New York: The Penguin Press.
-------- 1996b. "Comment on '"Many Minds' Interpretations of Quantum Mechanics" by Michael Lockwood", *British Journal of the Philosophy of Science* 47, pp. 222-8.
Devitt, M. 1990. "Transcendentalism about Content", *Pacific Philosophical Quarterly* 71, pp. 247-63.
DeWitt, B. S. 1971. "The Many-Universes Interpretation of Quantum Mechanics", reprinted in DeWitt and Graham 1973.
DeWitt, B. S., N. Graham, (eds.) 1973. *The Many-Worlds Interpretation of Quantum Mechanics*. Princeton: Princeton University Press.
Dieks, D. 2006. "Becoming, Relativity and Locality", in Dieks (ed.) *The Ontology of Spacetime*. Amsterdam: Elsevier.
Donald, M. J. 1999. "Progress in a Many-Minds Interpretation of Quantum Theory", *Los Alamos E-Print Archive,* url = http://www.arxiv.org/abs/quant-ph/9904001.
-------- 1997. "On Many-Minds Interpretations of Quantum Theory", *Los Alamos E-Print Archive,* url = http://www.arxiv.org/abs/quant-ph/9703008.
-------- 1995. "A Mathematical Characterization of the Physical Structure of Observers", *Foundations of Physics*, 25, pp. 529-71.
-------- 1990. "Quantum Theory and the Brain", *Proceedings of the Royal Society,* Series A, Vol. 427, pp. 43-93.
Dorato, M. 2002. "Kant, Gödel and Relativity", in. P. Gardenfors, K. Kiljania-Placek, J. Wolenski, eds. *Proceedings of the invited papers for the 11th International Congress of the Logic, Methodology and Philosophy of Science*. Synthese Library, Dordrecht: Kluwer, pp. 329-46.
Dretske, F. 1995. *Naturalizing the Mind*. Cambridge, Mass.: The MIT Press.
-------- 1993. "Conscious Experience", *Mind*, 102(406), pp. 263-83.
-------- 1981. *Knowledge and the Flow of Information*. Cambridge: Cambridge University Press.
Döring, K. 1972. *Die Megariker. Kommentierte Sammlung der Testimonien*. Amsterdam.
Earman, J. 1995. *Bangs, Crunches, Whimpers, and Shrieks: Singularities and Acausalities in Relativistic Spacetimes*. New York: Oxford University Press.
Earman, J., L. Ruetsche 2005. "Relativistic Invariance and Modal Interpretations", *Philosophy of Science*. Vol. 72, N. 4, pp. 557-83.
Eccles, J. 1994. *How the Self Controls Its Brain*. Berlin: Springer.

-------- 1990. "A unitary hypothesis of mind-brain interaction in the cerebral cortex", *Proceedings of the Royal Society of London* B240, pp. 433-51.
Einstein, A. 1905. "On the Electrodynamics of Moving Bodies", reprinted in *The Principle of Relativity*, New York: Dover Publications, 1952, pp. 35-65.
Einstein, A., Podolsky B., Rosen N. 1935. "Can Quantum-Mechanical Description of Reality be Considered Complete?", *Physical Review,* 47, pp. 777-80.
Everett, H. 1957. "'Relative State' Formulation of Quantum Mechanics", *Reviews of Modern Physics* Vol 29, 3, pp. 454-62.
Feigl, H. 1958. "The 'Mental' and the 'Physical'", *Minnesota Studies in the Philosophy of Science* 2, pp. 370-497.
Feigl, H., G. Maxwell, (eds.) 1962. *Minnesota Studies in the Philosophy of Science III*. Minneapolis: University of Minnesota Press.
Feynman, R. P. 1967. *The Character of Physical Law*. Cambridge, MA: MIT Press.
Feynman, R. P., Leighton, R. B., and Sands, M. 1963. *The Feynman Lectures on Physics, I*, New York: Addison-Wesley.
Field, H. 2003. "Causation in a Physical World", in M. Loux and D. Zimmerman, eds., *The Oxford Handbook of Metaphysics*, Oxford: Oxford University Press.
Fischer, J. M. 1995. *The Metaphysics of Free Will. An Essay on Control.* Oxford: Blackwell.
-------- 1988. "Freedom and Miracles". *Nous* 22, pp. 235-52.
-------- 1984. "Power over the Past", *Pacific Philosophical Quarterly*, 65, pp. 335-50.
Fischer, J. M., M. Ravizza 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.
Fitzgerald, G. F. 1889. "The Ether and the Earth's Atmosphere", *Science* 13, p. 390ff.
Flew, A. 1955. "The Third Maxim", *The Rationalist Annual,* 1955.
Fodor, J. A. 1994. "The Mind-Body Problem", in Warner and Szubuka (eds.) *The Mind-Body Problem*. Oxford: Blackwell.
-------- 1990. *A Theory of Content and Other Essays*. Cambridge MA: Bradford/MIT.
-------- 1987. *Psychosemantics*. Cambridge MA: MIT Press/A Bradford Book.
Foster, J. 1996. *The Immaterial Self*. London and New York: Routledge.

Frankfurt, H. G. 1988. *The Importance of What We Care About: Philosophical Essays*. Cambridge: Cambridge University Press.
-------- 1987. "Identification and Wholeheartedness", in *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, ed. Ferdinand D. Schoeman. New York: Cambridge University Press. Reprinted in and cited from Frankfurt 1988.
-------- 1971. "Freedom of the Will and the Concept of a Person". *Journal of Philosophy*, 68. Reprinted in and cited from Frankfurt 1988.
-------- 1969. "Alternate Possibilities and Moral Responsibility", *Journal of Philosophy*, Vol 66., No. 23, pp. 829-839.
Gardner, M. 1970. "Mathematical Games: The Fantastic Combinations of John Conway's New Solitaire Game 'Life'", *Scientific American,* 223, pp. 120-3.
Gell-Mann M., J. Hartle 1990. "Quantum Mechanics in the Light of Quantum Cosmology", in W. Zurek (ed.), *Complexity, Entropy and the Physics of Information*, pp. 425-458. Reading: Addison-Wesley.
Gellner, E. 1957. "Determinism and Validity", in *The Rationalist Annual 1957*.
Geroch, R. 1970. "Domain of Dependence". *Journal of Mathematical Physics* 11, pp 437-49.
Ghirardi, G. 2007. "Collapse Theories", in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy (Fall 2007 Edition)*, URL = http://plato.stanford.edu/archives/fall2007/entries/qm-collapse/.
Ghirardi, G. C., A. Rimini, T. Weber 1986. "Unified Dynamics for Microscopic and Macroscopic Systems". *Physical Review* D 34.
Gillett, C., B. Loewer (eds.) 2001. *Physicalism and Its Discontents*. New York: Cambridge University Press.
Ginet, C. 1990. *On Action*. New York: Cambridge University Press.
Goldman, A. I. 1979. "What Is Justified Belief?", in Pappas (ed.) 1979, pp. 1-23.
Goldstein, S. 2007. "Bohmian Mechanics", in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy (Fall 2007 Edition)*, URL = http://plato.stanford.edu/archives/fall2007/entries/qm-bohm/.
Gödel, K. 1990. *Collected Works*. Ed. by S. Feferman et al., Oxford: Oxford University Press.
-------- 1949a. "A Remark about the Relationship Between Relativity and Idealistic Philosophy," in Schilp 1949, pp. 557-62.

-------- 1949b. "An Example of a New Type of Cosmological Solutions of Einstein's Field Equations of Gravitation", *Review of Modern Physics* 21, pp. 447-50.
Graham, N. 1973. "The Measurement of Relative Frequency", in DeWitt and Graham 1973.
Griffiths, R. 1984. "Consistent Histories and the Interpretation of Quantum Mechanics", *Journal of Statistical Physics*, 36, pp. 219-72.
Grimm, R. H., D. D. Merill (eds.) 1988. *Contents of Thought*. Tucson: The University of Arizona Press.
Haag, Rudolf 1996, *Local Quantum Physics*. Berlin: Springer-Verlag.
Hajek, A. 1996. "Mises Redux-Redux: Fifteen Arguments against Finite Frequentism", *Erkenntnis* 45, pp. 209-27.
Heisenberg, W. 1958. "The Representation of Nature in Contemporary Physics", *Deadalus*, Summer 1958, pp. 95-108.
Hemmo, M, I. Pitowsky 2003. "Probability and Nonlocality in Many Minds Interpretations of Quantum Mechanics", *The British Journal for the Philosophy of Science*, 54(2), pp. 225-243. Preprint available online in the Los Alamos Archive.
Hobbes, T. 1962. *The English Works of Thomas Hobbes*. Vol 5. Ed. W. Molesworth. London: Scientia Aalen.
-------- 1958. *Leviathan*. Indianapolis: Bobbs-Merril.
Honderich, T. 2002. *How Free Are You?* Oxford: Oxford University Press.
Hume, D. 1975. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Ed. L. A. Selby-Bigge. Oxford: Clarendon Press.
James, W. 1890. *The Principles of Psychology*. New York: Henry Holt.
Kane, R., (ed.) 2002. *The Oxford Handbook of Free Will*. Oxford: Oxford University Press.
-------- 1996. *The Significance of Free Will*. New York, Oxford: Oxford University Press. Cited from the 1998 reprint.
-------- 1989. "Two Kinds of Incompatibilism", *Philosophy and Phenomenological Research* 219-54.
Kant, I. 1788/1949. *The Critique of Practical Reason, and Other Writings in Moral Philosophy*. Translated and edited by Lewis White Beck. Chicago: The University of Chicago Press.
-------- 1787/1970. *Critique of Pure Reason*. Translated by Norman Kemp Smith. Trowbridge&London: Macmillan St Martin's Press.

-------- 1783/1953. *Prolegomena to any Future Metaphysics that will be able to present itself as a Science.* Translated by Peter G. Lucas. Manchester: Manchester University Press.
Kapitan, T. 2002. "A Master Argument for Incompatibilism?" in Kane (ed.) 2002.
-------- 1996a. "Incompatibilism and Ambiguity in the Practical Modalities", *Analysis*, Vol. 56, No. 2, pp. 102-110.
-------- 1996b. "Modal Principles in the Metaphysics of Free Will", *Philosophical Perspectives* 10.
Kenny, A. 1975. *Will, Freedom and Power.* Oxford: Basil Blackwell.
Kim, J. 1993. *Supervenience and Mind.* Cambridge: Cambridge University Press.
Kirk, G. S., J. E. Raven, M. Schofield 1983. *The Presocratic Philosophers.* Cambridge: Cambridge University Press.
Klein, M. 1990. *Determinism, Blameworthiness and Deprivation.* Oxford: Clarendon Press.
Le Poidevin, R., ed. 1998. *Questions of Time and Tense.* Oxford: Oxford University Press.
Lehrer, K. 1968. "'Can's Without 'If's", *Analysis* 29.
--------, ed. 1966. *Freedom and Determinism.* New York: Random House.
Leibniz, G. 1952. *Theodicy; Essays on the Goodness of God, the Freedom of Man and the Origin of Evil,* Farrer, A. (ed.), Huggard, E.M. (trans.). New Haven: Yale University Press.
Lewis, C. S. 2002. "Religion without Dogma?", in Lesley Walmsley (ed.) *C. S. Lewis Essay Collection: Faith, Christianity and the Church*, London: HarperCollins.
-------- 1948. *Miracles.* London: Geoffrey Bles.
Lewis, D. 1990. "What Experience Teaches", in Lycan ed. 1990.
-------- 1986. *The Plurality of Worlds*, Oxford: Basil Blackwell.
-------- 1981. "Are We Free to Break the Laws?", *Theoria* 47, pp. 113-21.
Lockwood, M. 1996. "'Many Minds' Interpretations of Quantum Mechanics", *British Journal of the Philosophy of Science* 47, pp. 159-88.
Loewer, B. 1998. "Freedom from Physics: Quantum Mechanics and Free Will." *Philosophical Topics* 24.
-------- 1996. "Comment on Lockwood", *British Journal of the Philosophy of Science* 47, pp. 229-32.
Lorentz, H. A. 1892. "The relative motion of the earth and the ether", *Versl.Kon.Akad.Wetensch.* 1, p. 74ff.

Lovell, S. 2003. "C. S. Lewis's Case against Naturalism". Published on the Internet at www.csl-philosophy.co.uk.
Lucas, J. R. 2006. *Reason and Reality*. As yet, Mr Lucas has published the book only on his homepage, url: http://users.ox.ac.uk/~jrlucas/. He writes he may consider publishing it in a conventional format later.
-------- 1998. "Transcendental Tense II", *Aristotelian Society Supplementary Volume* 72, 29-43.
Lycan, W. ed. 1990. *Mind and Cognition: A Reader*. Oxford: Blackwell.
Malament, D. 1977. "Causal Theories of Time and the Conventionality of Simultaneity", *Nous* 11, pp. 293-300.
Markosian, N. 2004. "A Defense of Presentism", in Zimmerman 2004, pp. 47-82.
Maudlin, T. 1994. *Quantum Non-Locality and Relativity*. Oxford: Blackwell.
McGinn, C. 1996. "Can We Solve the Mind-Body Problem?", in *The Problems of Consciousness*. Oxford: Blackwell.
McTaggart, J. M. E. 1908. "The Unreality of Time", *Mind* 17, pp. 457ff.
Millikan, R. 1984. *Language, Thought and Other Biological Categories*. Cambridge MA: Bradford/MIT.
Mohyeldin Said, K. A., W. H. Newton-Smith, R. Viale, K. V. Wilkes, (eds.) 1990. *Modelling the Mind*. Oxford: Clarendon Press.
Montero, B. 2006. "What Does the Conservation of Energy Have to do with Physicalism?", *Dialectica* vol 60, no. 4, pp. 383-396.
Moore, G. E. 1912. "Free Will", in Moore, *Ethics*. Oxford: Oxford University Press.
Nozick, R. 1981. *Philosophical Explanations*. Cambridge MA: Harvard University Press.
O'Connor, T. 2000. *Persons and Causes. The Metaphysics of Free Will*. New York: Oxford University Press.
-------- ed. 1995. *Agents, Causes and Events. Essays on Indeterminism and Free Will*. New York, Oxford: Oxford University Press.
-------- 1993. "On the Transfer of Necessity", *Nous* 27, pp. 204-18.
Omnès, R. 1994. *The Interpretation of Quantum Mechanics*. Princeton: Princeton University Press.
Papineau, D. 2002. *Thinking about Consciousness*. Oxford: Oxford University Press.
-------- 2001. "The Rise of Physicalism", in Gillett and Loewer (2001).

-------- 1998. "Mind the Gap", in J. Tomberlin (ed.) *Philosophical Perspectives* 12.
Pappas, G. (ed.). *Justification and Knowledge: New Studies in Epistemology*. Dordrecht: Reidel.
Parfit, D. 1986. *Reasons and Persons*. Oxford: Oxford University Press.
Penrose, R. 1994. *Shadows of the Mind*. Oxford: Oxford University Press.
-------- 1989. *The Emperor's New Mind: Concerning Computers, Minds, and Laws of Physics*. Oxford: Oxford University Press.
-------- 1979. "Singularities and Time Asymmetry", in Hawking S. W. and W. Israel, eds. *General Relativity: an Einstein Centenary Survey*, Cambridge: Cambridge University Press.
Plantinga, A. 1994. *Naturalism Defeated*. Draft published on the Internet: www.homestead.com/philofreligion/files/alspaper.htm.
-------- 1993. *Warrant and Proper Function*. Oxford: Oxford University Press.
Plantinga, C., Jr. 1993. "Not The Way It's S'pposed to Be: A Breviary of Sin", *Theology Today*, Vol 50. No. 2, July 1993.
Popper, K. 1982. *Quantum Theory and the Schism in Physics*. Hutchison.
-------- 1959. "The Propensity Interpretation of Probability". *British Journal for the Philosophy of Science* 10, pp. 25-42.
Putnam, H. 1983. *Realism and Reason, Philosophical Papers Volume 3*. Cambridge: Cambridge University Press.
-------- 1975a. *Mathematics, Matter and Method, Philosophical Papers Volume 1*. Cambridge: Cambridge University Press.
-------- 1975b. *Mind, Language and Reality, Philosophical Paper Volume 2*. Cambridge: Cambridge University Press.
-------- 1967. "Time and Physical Geometry", *Journal of Philosophy* 64, pp. 240-7. Reprinted in Putnam 1975a.
-------- 1962. "The Analytic and the Synthetic", in Feigl and Maxwell (eds.) 1962, reprinted in Putnam 1975b, pp. 33-69.
Quine, W. v. O. 1987. *Quiddities*. Cambridge MA.: Harvard University Press.
-------- 1969. *Ontological Relativity and Other Essays*. New York: Columbia University Press.
Reichenbach, H. 1924. *Axiomatik der relativischen Raum-Zeit-Lehre*. Braunschweig: Friedrich Vieweg und Sohn.
Reid, T. 1895. *The Works of Thomas Reid, D. D.*, Eighth Edition, Ed. Sir William Hamilton, Edinburgh.

Rietdijk, C. W. 1966. "A Rigorous Proof of Determinism Derived from the Special Theory of Relativity", *Philosophy of Science*, 33, 341-4.

Rosenberg, J. 1986. "Intention and Action among the Macromolecules", in N. Rescher, ed., *Current Issues in Teleology*. Lanham, New York: University Presses of America.

Rubin, M. A. 2001. "Locality in the Everett Interpretation of Heisenberg-Picture Quantum Mechanics", published on the Internet, in the Los Alamos Archive, url = arxiv:quant-ph/0103079v2.

Russell, B. 1986. *The Collected Papers of Bertrand Russell*. London: George Allen and Unwin.

-------- 1972. *A History of Western Philosophy*, New York: Simon and Schuster.

-------- 1912-3. "On the Notion of Cause", *Proceedings of the Aristotelian Society*, 13, pp. 1-26.

Ryle, G. 1949. *The Concept of Mind*. London: Hutchinson and Co. (reprinted in 1958).

Saunders, J. T. 1968. "The Temptations of 'Powerlessness'", *American Philosophical Quarterly* 5, pp. 100-8.

Saunders, S. 2002. "How Relativity Contradicts Presentism", in C. Callender, ed., *Time, Reality, and Experience*. Cambridge: Cambridge University Press.

-------- 2000. "Tense and Indeterminateness", *Philosophy of Science*, 67, S600-611.

-------- 1998. "Time, Quantum Mechanics, and Probability", *Synthese*, 114, pp. 405-44.

-------- 1996. "Time, Quantum Mechanics, and Tense", *Synthese*, 107, pp. 19-53.

-------- 1995. "Time, Quantum Mechanics, and Decoherence", *Synthese*, 102, pp. 235-66.

Savitt, S. forthcoming. "The Transient *Nows*", in Myrvold, W. C., J. Christian, eds., *Quantum Reality, Relativistic Causality, and Closing the Epistemic Circle: Essays in Honour of Abner Shimony*. Springer, forthcoming. Available at Steven Savitt's homepage: http://www.philosophy.ubc.ca/faculty/savitt/Research/The%20Transient%20nows%20(Rev).pdf.

Schilp, P. (ed.) 1949. *Albert-Einstein: Philosopher-Scientist*. LaSalle, Illinois: Open Court.

Shimony, A. 1986. "Events and Processes in the Quantum World", in R. Penrose, C. J. Isham (eds.) *Quantum Concepts in Space and Time*, pp. 182-203. Oxford: Clarendon Press.
Shoemaker, S. 1969. "Time without Change", *The Journal of Philosophy*, 66, pp. 363-381.
Skinner, B. F. 1974. *About Behaviorism*. New York: Vintage.
-------- 1962/1948. *Walden Two*. New York: Macmillan.
-------- 1953. *Science and Human Behavior*. New York: Macmillan.
Slote, M. 1982. "Selective Necessity and the Free-Will Problem", *Journal of Philosophy* 79, pp. 5-24.
Smith, Q. 1998. "Absolute Simultaneity and the Infinity of Time", in Le Poidevin (ed.) 1998, pp. 135-183.
Stapp, H. 2007. "Quantum Mechanical Theories of Consciousness", in Velmans, M. and S. Schneider (eds.) 2007.
-------- 1993. *Mind, Matter and Quantum Mechanics*, New York: Springer.
Stein, H. 1991. "On Relativity Theory and Openness of the Future", *Philosophy of Science* 58, pp. 147-167.
-------- 1968. "On Einstein-Minkowski Space-Time". *Journal of Philosophy*, 65, pp. 5-23.
Stich, S., S. Laurence 1994. "Intentionality and Naturalism", in P. French, T. Uehling & H. Wettstein, (eds.) *Midwest Studies in Philosophy,* vol. 19: Naturalism, pp. 159-82. Notre Dame, Indiana: Notre Dame.
Strawson, G. 1986. *Freedom and Belief*. Oxford: Clarendon Press.
Strawson, P. F. 1962. "Freedom and Resentment", *Proceedings of the British Academy*, 48.
Tanney, J. 1995. "Why Reasons May Not Be Causes", *Mind and Language*, vol. 10, nos. 1-2, pp. 103-26.
Taylor, C., D. C. Dennett 2002. "Who's Afraid of Determinism? Rethinking Causes and Possibilities", in Kane (ed.) 2002.
Tegmark, M. 1998. "The Interpretation of Quantum Mechanics: Many Worlds or Many Worlds?", *Fortschr. Phys.* 46, pp. 855-62.
Tooley, M. 1997. *Time, Tense, and Causation*. Oxford: Clarendon Press.
Vaidman, L. 2000. "Many-Worlds Interpretation of Quantum Mechanics", in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy (Summer 2002 Edition)*, URL = http://plato.stanford.edu/archives/sum2002/entries/qm-manyworlds/.
van Fraasen, Bas C. 1991, *Quantum Mechanics: An Empiricist View*. Oxford: Oxford University Press.

van Inwagen, P. 2002. *Metaphysics*. Boulder: Westview Press.
-------- 1983. *An Essay on Free Will*. Oxford: Clarendon Press.
Velmans, M. and S. Schneider (eds.) 2007. *The Blackwell Companion to Consciousness*. Oxford: Blackwell.
Vihvelin, K. 1988. "The Modal Argument for Incompatibilism", *Philosophical Studies* 53.
von der Mühl, ed. 1922. *Epicurus*. Leipzig: Teubner.
von Mises, R. 1931. *Warscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik*. New York: Rosenberg.
von Neumann, J. 1955. *Mathematical Foundations of Quantum Mechanics*. Princeton: Princeton University Press.
Wallace, D. 2002. "Worlds in the Everett Interpretation", *Studies in the History and Philosophy of Modern Physics* 33, pp. 637-61.
Watson, G., ed. 1982. *Free Will*. Oxford: Oxford University Press.
-------- 1975. "Free Agency", *Journal of Philosophy* 72.
Weihs, G., T. Jennewien, C. Simon, H. Weinfurter, A. Zeilinger 1998. "Violation of Bell's Inequality under Strict Einstein Locality Condition", *Physical Review Letters*, 81, 5039-43.
Weinstein, W. 2002. "The Review of Palle Yourgrau's *Gödel Meets Einstein, Time Travel in the Gödel Universe*", *The Philosophical Review*, Vol. 111, No. 1. (January 2002), pp. 148-152.
Weyl, H. 1949. *Philosophy of Mathematics and Natural Science*. Princeton: Princeton University Press.
Wheeler, J. A., Zurek, W. H., eds. 1983. *Quantum Theory and Measurement*. Princeton: Princeton University Press.
Widerker, D. 1987. "On an Argument for Incompatibilism". *Analysis* 37, pp. 37-41.
Wiggins, D. 1998. "Towards a Reasonable Libertarianism", in *Needs, Values and Truth* (third edition). Oxford: Oxford University Press.
Wigner, E. P. 1997. *The Collected Works of Eugene Paul Wigner: Part A. The Scientific Papers III, Part I. Particles and Fields; Part II. Foundations of Quantum Mechanics*. Ed. By A. S. Wightman. Berlin.
-------- 1967. *Symmetries and Reflections*. Cambridge, MA.: MIT Press.
Wilks, Y., D. Partridge, (eds.) 1988. *Sourcebook on the Foundations of Artificial Intelligence*, New Mexico University Press.
Willard, D. 2000. "Knowledge and Naturalism", in *Naturalism, A Critical Analysis* ed. by W. L. Craig and J. P. Moreland, London: Routledge.

Wittgenstein, L. 1953/1967. *Philosophical Investigations*, trans. by G. E. M. Anscombe, third edition, 1967. Oxford: Blackwell.
Yourgrau, P. 1999. *Gödel Meets Einstein, Time Travel in the Gödel Universe*, Chicago: Open Court.
Zeh, H. D. 1970. "On the Interpretation of Measurement in Quantum Theory", *Foundations of Physics* 1, pp. 69-76.
Zimmerman, D. W. (ed.) 2004. *Oxford Studies in Metaphysics*. Oxford: Oxford University Press.
Zurek, W. H. 1991, "Decoherence and the Transition from Quantum to Classical", *Physics Today* 44, pp. 36-44.