(Metasemantically) Securing Free Will

Jason Turner

The Australasian Journal of Philosophy 91.2 (2013): 295–310

Abstract

Metasemantic security arguments aim to show, on metasemantic grounds, that even if we were to discover that determinism is true, that wouldn't give us reason to think that people never act freely. Flew's [1955] Paradigm Case Argument is one such argument; Heller's [1996] Putnamian argument is another. In this paper I introduce a third which uses a metasemantic picture on which meanings are settled as though by an ideal interpreter. Metasemantic security arguments are widely thought discredited by van Inwagen's [1983] Martian Manipulation objection. I argue that van Inwagen's objection, if right, can be parodied to undercut metasemantic arguments which aim to show that deliverances of physics do not tell us that no objects are solid. A diagnosis of where the parody objection breaks down against the Ideal Interpreter Argument arguments is then used to resist the objection as applied to that argument. I go on to defend the argument from the charge that it relies on a ham-fisted version of interpretivism.

Keywords: metasemantics, free will, determinism, paradigm case, interpretation, compatibilism

We care about free will, at least in part, because we hope that we have it. Free will is tied up with autonomy, moral responsibility, and our picture of ourselves as active contributors to the world we live in. Say that someone *has free will* if she sometimes acts freely, and let the *free will thesis* be the thesis that some people have free will. If it is true, all the better; if not, all the worse.

Determinism, writ large, is the thesis that the remote past plus the laws of nature fix absolutely everything else. We care about determinism, at least in part, because we worry that it might rule out free will. Our best physical theories (at least until the 1930's or so) were deterministic, and coupled with a robust physicalism led to determinism writ large. If we can't be both determined and free, then our freedom might at best appear hostage to the ultimate deliverances of physics.

Two theses are compatible if and only if their conjunction is possible. *Compatibilism* says that the free will thesis is compatible with determinism. (In short: free will is compatible with determinism). We care about compatibilism and its denial, incompatibilism, partly because its truth bears on whether determinism threatens free will. And we care about this question partly because we care about whether physics might someday give us reason to think we're not free.

Call a thesis *p secure from* a thesis *q* when discovering the truth of *q* wouldn't give us a reason to doubt that *p*. On the above way of thinking, we care about compatibilism partly because we care about whether the free will thesis is secure from determinism. (In short: whether free will is secure from determinism).

We might care about compatibility for other reasons. Even if completely confident that we have free will, we might think questions of compatibility help us better understand its nature. But security is clearly *a* reason to care about compatibility, and one that has animated a number of authors.

One style of argument for the security of free will from determinism is *metasemantic*: it hinges on details of the meaning of 'free' and the way this meaning has been fixed.[1] Antony Flew's famous Paradigm Case Argument [1955] is one such argument, Mark Heller [1996] has given another, and later in the paper I will present a third.

Most philosophers think that metasemantic security arguments aren't worth bothering with, having been soundly refuted by Peter van Inwagen [1983: 106–13]. I disagree: at least some such arguments can adequately respond to van Inwagen's criticisms. After describing Flew's Paradigm Case Argument (§1), I present a new security argument based on a certain interpretivist metasemantics (§2). After a brief aside (§3), I defend the new argument from van Inwagen's criticism (§5), and then go on to defend the interpretivist argument from a further objection (§6). My aim here isn't to offer a full-fledged defence of the metasemantic security of free will from determinism. That would require settling which metasemantic theory is correct, which I cannot hope to do here. Rather, I hope to show that, despite philosophers' usual attitude towards them, metasemantic security arguments look promising and deserve further attention.

1 The Paradigm Case Argument

Flew's Paradigm Case Argument was an argument for compatibilism, but it's useful to think of it as a security argument with compatibility an added bonus stemming from the argument's background metasemantic theory. In particular, the argument hinges on the *molecular* picture of language dominant in Flew's time. This picture weds expressions' meanings to how they were learned. We teach words in only two ways, it was thought: by ostend-

[1] More carefully, it hinges on the details of what philosophers are pleased to call the 'semantic value' of 'free': the truth-conditional contribution 'free' makes to sentences it appears in, stripped of any further pragmatic contribution.

ing — pointing to a green thing, for instance, and saying '"green" applies to things of *that* color' — or by defining using already-understood terms.

Call the words we learn only by ostension *atomic*. Meanings of atomic terms were thought fixed by the cases to which one had to ostend to teach them, so that the terms ended up applying to all and only cases relevantly like those paradigms. The *molecular* expressions, on the other hand, could be learned through definitions. Their meanings were thought of as the complexes of atomic expressions that resulted from unpacking the relevant definitions to the atomic level.

The Paradigm Case Argument — or, at least, the security-relevant part of it — runs:

> 'Free' is not a molecular expression — 'we are not dealing with a compound descriptive expression correctly formed of words which can and have been given sense independently' [Flew 1955: 151] — and therefore is an atomic one. But since the meaning of an atomic expression is fixed by the cases to which we ostend when teaching it — the 'paradigm cases' — the meaning of 'free' must also be so fixed. So 'free' applies to some actions (at a minumum, the paradigm cases themselves). If we discover that determinism is true, then we learn that these paradigms are determined; but they are still similar to themselves, so they still satisfy 'free'. So if we discovered the truth of determinism, that would not give us reason to think no actions satisfy 'free', and thus no reason to think that no actions are free.[2]

Motivation for the molecular picture gets us from here to compatibility. The picture was driven, in part, by epistemological concerns. Philosophers wondered how competent language users knew when to use an expression. The molecular picture held that these users knew when to use a term thanks to how they learned it: speakers knew to use an atomic expression when they encountered a case relevantly like its paradigms, and to use a molecular expression when they saw that its definition was satisfied. By insisting that the meaning of an expression corresponded closely to how it was learned, philosophers guaranteed that language-users who used expressions in this way were bound to use them correctly.

To make this work, determining whether a case was relevantly like the paradigms would have to be something any competent language user could

[2] Operative in this argument and throughout the paper is the premise that an action satisfies 'free' iff it is free (and similarly for other terms), which lets us move quickly between talk of terms and talk of free actions. Since the premise isn't necessarily true, these moves can be dangerous, but throughout the premise is used only in modal contexts where the worlds are to be 'considered as actual' [Yablo 2002], and so the move is safe. For brevity, I won't mention this move again.

do. Thus, features that would not normally be epistemically available to a competent language user couldn't be relevant to an atomic expression's meaning. A speaker competent with an atomic expression α had to be able to tell just by looking whether or not α applies in a given situation. So anything that cannot be determined of a situation just by looking could not be relevant to the application of α.

We cannot tell whether determinism is true just by looking. The paradigm cases would look the same either way. Thus (goes the line of thought), whatever meaning 'free' has would be the same whether or not determinism were true. And since, *if* determinism were true, 'free' would have a meaning compatible it, then 'free' *does* have such a meaning, regardless of determinism's truth.

2 The Ideal Interpreter Argument

Unfortunately for the Paradigm Case Argument, the molecular picture of language it rests on has been widely discredited. Other metasemantic security arguments can be given, though.

Mark Heller [1996] gives one stemming from the metasemantic picture of natural kinds developed by Saul Kripke [1972] and Hilary Putnam [1975]. On that picture, natural kind terms are (to a first approximation) introduced by an 'initial baptism' — someone pointing at something and says 'let stuff like that be called α' — and their meanings are fixed by the deep explanatory structure of the stuff pointed to. Heller suggests that 'free action' might be a natural kind term, and that if it is, and if determinism is true, then whatever action pointed at in the initial baptism was a determined one. Thus, if 'determinism' is true, 'free action' applies to at least some determined actions (the ones used for baptism, at a minimum) and so some actions are free.

Heller's argument relies on 'free' being a natural kind term, which some have objected to. [Daw and Alter 2001: 349–50] But a different argument, which needs no such claim, can also be given. This argument relies on an *interpretivist* metasemantic picture. This picture starts from the thought that meaning is fixed by use. More precisely, a linguistic community's use of words is embedded in a complex pattern of behavior and response to environment. A word's meaning is fixed by the way it is embedded into this pattern.[3] We get a grip on the way use fixes meaning by pretending meanings get assigned by an 'ideal interpreter'. This interpreter is ideal in two ways: she lacks cognitive limitations, and has access to all non-semantic facts about the community she interprets. She knows how the community uses their terms, how they behave when those terms are used, how they are dis-

[3] David Lewis [1983a] argues that we need to add a metaphysical element to meaning determination on this sort of picture; we'll return to this in §6.1.

posed to behave in various counterfactual circumstances (which may or may not involve the use of other terms), and so on. And she uses this information to decide what meaning to give each expression. [Lewis 1974]

Her decision is constrained by various principles. She should interpret to some degree *compositionally*, for instance: meanings assigned to complex expressions should be a function of the meanings assigned to their parts. Crucially for our purposes, she should interpret *charitably*: meanings assigned should maximize the truthfulness, rationality, and understandability of the community in question.

From this perspective, we can argue:

> Ordinary speakers confidently and unhesitatingly call some actions 'free'. Because ordinary speakers are confident in these ascriptions, and these ascriptions are so widespread, charity pressures an ideal interpreter to assign 'free' a meaning that applies to these actions. And this pressure occurs whether or not determinism is true: the interpreter is supposed to make us understandable, not do philosophy. So if the ideal interpreter finds herself in a deterministic universe, she will try to assign 'free' a meaning satisfied by, at a minimum, the actions to which ordinary speakers confidently apply it. Thus, if we discover determinism is true, we gain reason to think the ideal interpreter gave 'free' a meaning satisfied by some determined actions.

Notice that, unlike the Paradigm Case Argument, this argument does not directly support compatibilism. If indeterminism is true, the ideal interpreter may have given 'free' a meaning that cannot apply to determined actions.[4]

This argument is less straightforward than the others. It works only if the pressure on the ideal interpreter to interpret 'free' charitably isn't outweighed by other pressures. But there are other pressures. At least some people are prone to assert 'If determinism is true, then no one has ever acted freely.' Even in a deterministic world, if she doesn't interpret their other words bizarrely, she can make these assertions true only if she keeps 'free' from being satisfied.

This pressure should not outweigh the pressure from charity, though, and for three reasons. First, although some speakers — mainly philosophers — *do* assert this, others assert its negation. The ideal interpreter can't make *everyone* right. One group of assertions pushes her to give 'free' an incompatibilist meaning; the other group pushes her the other way. Plausibly, these pressures cancel out, and the tie will be broken by ordinary speakers' confident applications of 'free' to cases.

[4] The same goes for Heller's argument, as he points out [1996: n. 8]. In both cases this is because the meaning the metasemantics predicts for 'free' if determinism is false may be different from the one it predicts if determinism is true.

Second, ideal interpreters plausibly owe less to high-falutin' theoretical claims than to confident first-order judgments. Ordinary, confident deployment of 'free' should get more weight than philosophical speculation.

This idea isn't new. We can understand Moore's anti-skeptical response as insisting that an ideal interpreter is under more pressure to assign meanings in a way that validates 'I know that I have hands' — a paradigm case of knowledge — than to validate the philosophical theses that entail skepticism. There is something intuitively right about this. The ideal interpreter's job is to make as much sense of us as possible — to make us rational and explain how we successfully navigate our environs. If even paradigm uses of 'knows' are in error, then — given the way epistemological concerns infect so much of what we do — our behavior will be hard to understand. And since concerns of freedom are as important and as infectious as epistemic concerns, similar remarks apply to the ideal interpreter's treatment of 'free'.

Finally, an interpreter is supposed to be charitable to an *entire community*, not simply to a handful of 'enlightened' speakers within that community. Philosophical pronouncements about freedom are far fewer than confident deployments of 'free'. Weight of numbers suggests that it is most charitable to interpret 'free' as applying to at least some actions — even if we happen to be in a deterministic world.

3 An Aside: Words and Concepts

The arguments we've considered suppose that 'free' was an ordinary term of natural language — part of our linguistic endowment, like 'blue', 'water', or 'justice' — so that a standard metasemantic story applies to it. But some may think, with van Inwagen [1989: 400], that it is a philosophical term of art. If so, mightn't a different metasemantic story apply?

I see little reason to worry. Perhaps 'free', as used by philosophers when discussing the sort of freedom relevant to moral responsibility, is a term of art. I doubt the concept it stands for is. I suspect rather that, long before philosophical theorizing, we have the relevant concept of *freedom* as part of our cognitive endowment. Children may deploy this concept when they complain to their parents that 'it's not my fault!' or 'I couldn't help it!', or while seething about a friend's errant behavior. If 'free' is introduced as a term of art, it is a term designed to track a concept we already have and deploy.

Some evidence for this comes from the relative ease with which laypeople pick up 'free' and incorporate it into their natural vocabulary. In this, it is little like paradigm terms of art such as 'instantiation', 'satisficing', or 'supervenience'.

More evidence comes from the possibility of disagreement over potential definitions of the term. If 'free' were a purely technical term of art, its content would be entirely fixed by stipulative definition. Authors who define it differently would simply talk past each other, and people would not be competent with the term until they had mastered the definition.

We find the opposite, though. 'Free' is remarkably easy to pick up and develop confident first-order judgments with even in the absence of a definition.[5] And philosophers who disagree about the definition generally take themselves to be having a substantive, non-verbal disagreement about the *right* way to define 'free'.

This suggests the term of art, if such it is, stands for a pre-theoretical concept. But concepts are just the words of the mind, every bit as contentful as words of English. And each metasemantic story told above can be fitted to apply to concepts rather than words, which means each argument above can be re-tooled so as to be about the concept *free* rather than the word 'free'. If the concept *free* is part of our natural cognitive endowment, these arguments can stand as proxy for the ones presented above.

For continuity, I will continue to focus on the language-centered versions of the arguments; concerned readers can interpret me as covertly talking about concepts instead.

Finally, to stave off confusion: distinguish *concepts* — our mental vocabulary, the basic unit of cognition — from *conceptions*. Our concept *cat* is the vehicle of our thoughts about cats; our conception of cathood is what we think cats are like. The two can come apart; as Putnam [1962] suggests, if all catlike creatures turn out to be robots controlled by Martians, then our concept *cat* covers these robots, but our conception will be radically mistaken. Philosophers sometimes use 'concept' for what I am calling a 'conception', but I won't follow them here.

4 Van Inwagen's Criticism

Peter van Inwagen [1983: 106–13] criticizes the Paradigm Case Argument as follows. First, he presents the thesis of *martian manipulation*:

(M) Whenever any human is (or ever has been) born, martians implant a tiny device in its brain that lets them control its every thought and action. Then, throughout the rest of that human's life, the martians control it from afar.

He then argues that, if 'free' gets its meaning the way the Paradigm Case Argument says, then even if we discover that martian manipulation is true, that discovery would not give us any reason to think that no actions satisfy 'free'. If martian manipulation is true, the paradigms themselves are

[5] Witness e.g. the surprisingly stable responses of participants in experimental work on lay intuitions about free will [e.g. Nahmias *et al.* 2005; 2006; Nichols and Knobe 2007] — participants with no previous definition of 'free' to underwrite their competence.

martian-manipulated actions. But the initial paradigms are sufficiently like themselves to satisfy 'free'; thus some actions, at least, will satisfy 'free' despite being controlled by martians.

Clearly, though, the discovery of martian manipulation *should* make us think that nobody acts (or ever has acted) freely. So, van Inwagen concludes, there must be something wrong with the martian manipulation argument, and by extension, with the Paradigm Case Argument it parodies.

Daw and Alter [2001: 350–2] argue that van Inwagen's criticism applies to Heller's argument, and I'm often told (in conversation) that it likewise undercuts the Ideal Interpreter Argument. Here's how the complaint goes against the Ideal Interpreter Argument: were it right, then if the actions we confidently call 'free' were manipulated by martians, the ideal interpreter would assign 'free' a meaning that applied to manipulated actions, and discovering this would give us no reason to think people aren't free. Since such a discovery *would* give us reason to think that people aren't free, the Ideal Interpreter Argument must be in error.

5 Responding to van Inwagen's Criticism

I think the Ideal Interpreter Argument can (with auxiliary premises) get around the problem. The response is best illustrated by a parallel case.

5.1 Solidity: A Warm-up Exercise

Most people pre-theoretically assent to:

(1) Something is solid only if there is no empty space in it.

But modern physics has taught us that the macroscopic objects we think of as solid — chairs, stones, and other medium-sized dry goods — are composed of swarms of microphysical particles held rigidly together by various forces. Many of these particles are tremendously far from each other, in proportion to their sizes; as a result, these medium-sized goods are largely made up of empty space.[6] Did this discovery teach us that there are no (macroscopic) solid objects?

Here's an argument, based on the interpretivist metasemantic picture, that it did not:

[6] At least, I'm often told modern science has taught us this. The picture seems to suppose that subatomic particles are like little billiard balls. If instead, as some interpretations of quantum mechanics have it, these particles are superposed throughout a comparatively large region, then I don't know how to make sense of the claim that the particles are 'tremendously far from each other, in proportion to their sizes'. At any rate, it doesn't matter here; we can construe the warm-up exercise as a thought experiment premised on subatomic particles in fact being like little billiard balls.

> Suppose meanings are assigned by an ideal interpreter, and suppose she has already assigned meanings to everything but 'solid', which she is working on now. Since we tend to assent to (1), she will be under some pressure to assign 'solid' a meaning that cannot be satisfied by things having empty space in them. But since we confidently apply it to tables, bricks, walls, and so on, which have empty space in them, she is under pressure to assign 'solid' a meaning that can be satisfied by things with empty space in them. Since our confident applications of 'solid' outweigh our folk conception of solidity (as manifest in e.g. (1)), the meaning of 'solid' — and thus, solidity — is compatible with things being filled with lots of empty space.

Many philosophers will find some variant of this argument compelling. But here is an objection that threatens any such argument. Consider the thesis of *universal hollowness*:

(UH) Every macroscopic object is hollow — a thin shell surrounding a vacuum. Whenever we cut these objects open, a deceitful demon automatically creates a new thin surface along the cut to cover what would otherwise be an opening into the object's hollow interior. We thus never realize that all objects are hollow.

We can parody the above argument and conclude that, even if we discover that universal hollowness is true, that will not give us reason to think that macroscopic objects are not solid. But since a discovery of universal hollowness should make us think *just that*, there must be something wrong with the parody argument, and by extension, the original.

5.2 *Solidity Regained*

Solidity is not in such a perilous state, though; the cases aren't as symmetrical as the objection supposes. The ideal interpreter is constrained by our confident deployment of terms, such as 'solid'. She is also constrained by our expressions of our *conception* of solidity. Thesis (1) expresses part of this conception. Other parts involve matter and mass-density distributions throughout solid objects, the (im)possibility of nesting solid objects, and so on.

Charity would have our ideal interpreter try to make us right both about our conception of solidity and our confident deployment of 'solid'. If our confident deployments of 'solid' are always in error, we will be hard to understand. But making us right about these deployments at the cost of making us *radically* mistaken about the nature of solidity also makes us hard to understand. If the ideal interpreter can make us right about the deployments without pulling the meaning of 'solid' *too far away* from our conception of solidity, she will do that; if not, she will defer to our conception and make the objects not satisfy 'solid'.

In the actual case, although objects have empty space in them, the complex interplay of forces, even distribution of mass, etc. of bricks, tables, etc. means the empty space in these objects acts very much the way we pre-theoretically expected of only completely filled space. Arguably, this preserves enough of our pre-theoretical conception of solidity for an ideal interpreter to assign 'solid' a meaning applying to these objects without doing too much damage to our conception. So she does.

If instead universal hollowness is true, bricks, tables, etc. are very unlike our conception of solidity; the interpreter can have 'solid' apply to these objects only by doing great violence to that conception. So she does not. The interpretivist can thus accept the failure of the universal-hollowness argument without giving up her argument for the security of solidity from the empty space actually in objects.

5.3 *Free Will Regained?*

Given the above, we can see that the Ideal Interpreter Argument emerges unscathed if the following hold: (1) Being manipulated by martians damages our conception of freedom enough that, ideal interpreters or initial baptisms notwithstanding (and bullet bitings aside), 'free' couldn't mean *that*. (2) Even if determinism pulls us away from our intuitive conception of free acts, it doesn't pull us *too far* away for an ideal interpreter to ever assign a meaning compatible with determinism to 'free'.[7]

Doubters will worry determinism *does* pull us too far away. Whether it does depends on what our conception of free acts is like, and how heavily we value various components of this conception. So metasemantic security arguments won't license an end-run around conceptual analysis. But the bar for security fans is lower than for those who want to establish compatibilism through conceptual analysis. For example, it may be that our conception of freedom has it that we freely act at a time only if we had a choice about any propositions which entailed that we so acted. Call this the *closure requirement*. The Consequence Argument [van Inwagen 1983: 56] shows that no determined action can satisfy the closure requirement. If an action has to satisfy the requirement to be free, then determined actions can't be free.

[7] Heller can, with a bit of work, respond to van Inwagen's criticism in a similar way. The trick is to insist that, in response to the '*qua* problem' [Devitt 1991: 436; 1981: 60-4; Sterelny 1983: 120], initial baptisms involve both an ostension and a conception of the ostended stuff, with the ultimate meaning of the baptized term fixed by a balance between the conception and the underlying explanatory properties of what's been ostended. If the underlying property is not 'too far away' from the conception, all goes according to plan; but when the underlying property *is* too far away (as it might be when what's ostended is universally hollow or manipulated by martians) the attempted baptism fails.

The traditional compatibilist response tries, more or less, to show that this closure requirement isn't part of our conception after all. But the security fan can grant that it *is* part of our conception, while insisting that it's not central enough to the conception to keep the ideal interpreter from ever assigning a meaning to 'free' that violates it.

The defender of the Ideal Interpreter Argument will want to show that, even if our conception has incompatibilist elements, some deterministic surrogates are 'close enough' to our conception for an ideal interpreter in a deterministic world to assign them to 'free'. To make good on this demand, we need a way to measure the 'distance' between potential meanings, on the one hand, and conceptions, on the other. Very roughly, the thought is that a potential meaning m_1 is closer than another m_2 to a conception of freedom if, thanks to the conception and setting aside other facets of 'free's use, the ideal interpreter would prefer to make m_1 rather than m_2 'free's meaning. (Note: the ideal interpreter might prefer m_1 to m_2 while still preferring some third option over both of them.)

Since I have not set for myself the task of establishing free will's security from determinism, I don't feel obliged to argue here that some meanings compatible with determinism are close enough to our conception for the ideal interpreter to potentially assign them. But there are reasons for optimism. The ideal interpreter's job is to make as much sense of us as she can. To do that, she will consider not just what our conception in fact says, but why it says it. Our concepts, and the terms that stand for them, play various important roles in our cognitive and social economy, and our conceptions of them serve to help them properly fill those roles. Insofar as assigning a meaning to a word or concept interferes with how we use it to interact with the world — the more it makes our interactions inappropriate, etc. — the more the ideal interpreter will resist such an assignment. And the more an assignment preserves of a concept's or term's socio-cognitive functional role, the more willing the ideal interpreter to make it.

There are reasons to think that the most central parts of our conception — the parts most important to the role it plays in our socio-cognitive economy — won't require incompatibilism. The conception grew out of social interactions with others that were sensitive to a number of factors — factors that compatibilists tend to highlight, such as whether agents 'identified' with their actions [Frankfurt 1971], or were properly responsive to reasons [Fischer and Ravizza 1998], or were coerced or otherwise manipulated into doing what they did. It evolved to help us decide when to target a given agent or action for the various reactive attitudes, and when to withhold such attitudes [Strawson 1962]. But the social interactions our conception is designed to handle don't seem sensitive to whether the microscopic state of the world fixed everything else. When we engage in the interactions, microphysics isn't on our radar. And it's not clear what useful purpose would be served by putting it there. (How does society benefit if two mentally and macrophysically duplicate agents, who differ only in that one of them is microsopically deterministic and the other isn't, are treated differently?) So, while we should expect the ideal interpreter to resist assigning 'free' a meaning applying to addicts, the manipulated, and those not responsive to reasons, we should not likewise expect her to resist a meaning just because it applies to determined agents.

A worry: If we grant that the ideal interpreter will resist giving 'free' a meaning applying to manipulated agents, do we give up the game? Several authors (e.g., Robert Kane [1996: 64–71] and Derk Pereboom [2001: 110–25]) have argued that manipulation is not significantly different than determinism. So if manipulation damages our conception, then doesn't determinism, too?

I doubt it. The manipulation arguments (very roughly) run:

(i) An agent manipulated to *A* in circumstances *C* is not free.

(ii) An agent manipulated to *A* in circumstances *C* is not significantly different (in any free-will-relevant respects) to one causally determined to *A* in *C*.

(iii) Therefore an agent causally determined to *A* in circumstances *C* is not free.[8]

As it stands, this is no argument against security: the friend of security can (but need not) consistently accept (iii) and also accept that the discovery of determinism wouldn't give us reason to think agents are unfree. (He would have to think the world is indeterministic to do so.[9])

But the worrier is thinking that, if meanings applying to manipulated agents are too far away from our conception, then ones applying to determined agents are, too. The comparable argument for that would have to run:

(i′) An ideal agent will not assign 'free' a meaning that applies to an agent manipulated to *A* in circumstances *C*.

(ii′) An agent manipulated to *A* in circumstances *C* is not significantly different (in any respect relevant to how our conception of freedom affects an ideal interpreter) to one causally determined to *A* in *C*.

[8] See McKenna [2008: 143] for a similar, but more sophisticated, presentation.

[9] If determinism is true and free will is secure from it, then free will is also compatible with it. So the friend of security will have to say that, if determinism is true, the manipulation argument fails — and she owes us a story as to where it does so. But the story isn't hard to come by: she will think that (ii) is false, but that it *seems* true because it's a (relatively minor) part of our conception of freedom. Its falsity will be like that of (1): a surprise we get thanks to a mismatch between the world and our conception of it.

(iii$'$) Therefore ideal agent will not assign 'free' a meaning that applies to an agent determined to A in C.

If successful, this argument would undercut metasemantic security.

I doubt premise (ii$'$). First: (ii) is supported by our finding *no principled reason* to rule the determined actions free and undetermined ones unfree. But we can't support (ii$'$) in the same way. There is a principled reason for the ideal interpreter, in light of our conception of freedom, to treat manipulated actions different than ordinary determined actions. Our conception came into being, and evolved to handle, interactions where manipulation might be an issue; it didn't come into being, and didn't evolve to handle, interactions where causal determinism might be an issue. Since she is concerned primarily with doing justice to interactions that our conception evolved to handle, she has a principled reason to care more about manipulation than determination.

Another stab for (ii$'$): The objector might claim that it's built into our conception of freedom that only such-and-so factors can make a difference to whether or not someone is free, where those factors don't in fact divide cases of manipulation from cases of determinism.

If so, then our conception has a 'no-difference' requirement as well as a 'no-manipulation' requirement. But this doesn't help: the ideal interpreter might be just as willing to forgo the no-difference requirement on our conception of freedom as she is to forgo the no-empty-space requirement on our conception of solidity. To underwrite (ii$'$), the no-difference requirement would have to say, inter alia, that the mere fact that one agent was manipulated whereas another wasn't cannot on its own make the first unfree and the second free. Given the social importance of sussing out manipulation, I'm skeptical that this is part of our conception at all; but even if it is, I can see no reason to think that it's so important to our conception's purpose that no ideal interpreter would dare violate it.

Assigning 'free' a meaning that applies to determined actions but not manipulated ones might do some violence to our conception. But it wouldn't follow that it does *enough* violence to prohibit such an assignment. To undermine the security argument, the mismatch between meaning and conception must be *severe enough* that an ideal interpreter wouldn't make such an assignment. Even if our conception lists certain factors as being the only things relevant to it, it's by no means clear — or even very plausible — that this list of factors is so central to the conception that no interpreter would dare violate it.

6 Another Objection to the Ideal Interpreter Argument

Interpretative charity is a complicated business, balancing a community's utterances, behaviors, dispositions, and so on to best make sense of people. Some readers might fear the Ideal Interpreter Argument depends on expecting the ideal interpreter to follow a ham-fisted 'make what people say true, come what may' policy, and that the argument will not go through if our interpreter is more sophisticated.

To see the objection, we need to see how the metasemantic picture might be complicated. A just-so story will illustrate ham-fisted charity's inadequacy. We will then consider two ways to improve our interpreter's policies, and consider their potential implications for the argument.

First, the just-so story: A long time ago, people said

(2) Whales are fish.

Later, they discovered several differences between whales and other animals they applied 'fish' to, and started saying instead

(3) When we said whales were fish, we were wrong.

If this just-so story is right, how should the ideal interpreter react?

The Ideal Interpreter Argument gave special consideration to ordinary, confident applications of a term to an object. In homage to the Paradigm Case Argument, let's call these confident applications *paradigms*; the Ideal Interpreter Argument assumes that an ideal interpreters should try to make paradigm applications true.

But that policy seems to lead her astray here. Since many people said (2), and said it with confidence, etc., the policy would have her make (2) true and (3) false. But (3) should be true and (2) false. We need to give the ideal interpreter instructions that will let her sometimes make paradigm applications false.[10]

6.1 Metaphysically Special Properties

One such instruction is distinctively metaphysical. Lewis [1984; 1983a: 45–55] has argued that an elite few properties are metaphysically privileged. These are the properties that make for objective similarity, figure in the laws of nature, and ground intrinsicality. And he has argued also that we should instruct ideal interpreters to interpret in a way sensitive to these properties.

The suggestion can be implemented in several ways, but a standard way appeals to some properties being more *elite* than others. (More elite properties are less gerrymandered than less elite ones.) The ideal interpreter is

[10] See Weatherson [2003] for a nuanced discussion of this sort of problem and some of the issues touched on below.

then instructed to interpret charitably while giving terms as elite a meaning as possible — to interpret so as to best blend charity and eliteness. When the two come apart, she must choose, and either can be sacrificed for the other when gains outweigh costs.

From this perspective, we might think that the ideal interpreter makes (2) false and (3) true for reasons of eliteness. Since all fish are more like each other than they are like whales (thanks to biological similarities and so forth), a meaning for 'fish' that applied equally to whales would be more gerrymandered than one which applied only to the fish. The ideal interpreter, sensitive to this fact, assigns 'fish' the more elite, whale-excluding meaning, making (2) false.

Here is an eliteness-inspired worry for the Ideal Interpreter Argument: even if the paradigms of 'free' are determined, if the most elite property in the vicinity is incompatible with determinism, won't the pressure to assign elite properties push the interpreter to give 'free' a meaning incompatible with determinism? If so, then wouldn't a discovery of the truth of determinism give us reason to think no actions satisfy 'free' after all?

I see no reason to think that the most elite property in the neighborhood of our use would be incompatible with determinism. Compatibilists and incompatibilists generally agree to a certain set of necessary conditions for free will, with the latter thinking a further condition — one which entails indeterminism, as it happens — is also necessary. If this is right, then the property the incompatibilist associates with 'free' seems a gerrymander of the one the compatibilist thinks it means plus a further condition, and thus less elite.

But even if the determinism-excluding property is more elite, the worry is misguided. The interpreter is supposed to find an interpretation that *best blends* charity and eliteness. An interpretation that makes 'fish' apply to many paradigms tips its hat to charity; one which refrains applying it to all paradigms on biological grounds tips its hat to eliteness. But one that makes 'fish' apply to *no* paradigms jettisons charity entirely, no matter how elite the property it assigns to 'fish'. Likewise, an interpretation that makes 'free' apply to some but not all paradigms tips its hat to charity; one that makes it apply to no paradigms at all jettisons charity entirely. This is hardly the 'best blend' of charity and eliteness.

6.2 Dispositions to Retract

Another diagnosis of the fish/whale case — one I prefer — appeals to an expanded conception of charity.[11] The idea here is, roughly, that the ideal interpreter is charitable not only to what we in fact say, but also to what

[11] My preference is not founded on any widespread dismissal of eliteness in metasemantics, but rather on particular concerns about how the eliteness idea should be implemented.

we are disposed to say in various counterfactual circumstances. Facts about how we 'use' a term should be construed to include facts about how we are disposed to use a term in various situations; the fact that we are disposed to use a term in a certain situation should count as part of how we use it, even if that situation never actually arises and so we never manifest the disposition.

For instance, even when we were uttering (2), we were disposed to utter (3) upon learning certain biological facts. It turned out that we eventually learned those facts, and manifested our dispositions by our coming to say (3). If the ideal interpreter wants to be charitable not just to those who utter (2), but to *all* the speakers of the language at any time whatsoever, then she will have at least some reason to interpret (3) as true — which in turn gives her some reason to interpret (2) as false. Since the move to reject (2) and accept (3) is based on the community's bettering its epistemic state, and since even those who never utter (3) *would* have if they had lived long enough, the (3)-favoring interpretation looks *more* charitable than the (2)-favoring one.

Even if humanity had been annihilated before learning the relevant biology, the dispositions would have remained. If the ideal interpreter is to be charitable not just to our actual selves but also to our possible selves who manifest these dispositions, she will have a reason to interpret 'fish' so as to exclude whales and makes (3) true. In short: even if we in fact apply a term α to a certain class of cases C, *if we are disposed to retract these claims upon coming to be better informed*, that gives the ideal interpreter reason to *not* interpret α as applying to the cases in C.

This has the makings for a powerful attack on the Ideal Interpreter Argument. Even if we apply 'free' to a lot of actions, if we are disposed to retract these ascriptions upon learning the truth of determinism, and if determinism is true, then the ideal interpreter will have a reason to assign 'free' a meaning that is not compatible with determinism (and doesn't apply to any actions at all).

The attack is potentially devestating. But successfully launching it requires establishing an empirical premise about our dispositions: that we are disposed to call all actions unfree upon learning the truth of determinism.

The truth of this empirical claim is unclear, and we in fact have some reason to doubt it. Nahmias *et al.* [2005; 2006] presented non-philosophers with vignettes describing a deterministic universe, and then asked them to answer whether or not the agents acting in the vignettes acted freely. Across a variety of conditions, including a variety of descriptions of determinism, about two-third of respondents answered that the agents did act freely, and only one-third answered that they did not. Presumably, though, anyone who calls an agent free after being given a clear description of how they are determined[12] isn't disposed to stop calling people free after learning the truth

[12] In one vignette, respondents are told that a supercomputer can predict from the state of the universe in the remote past and the laws of nature, with 100% accuracy and long before

of determinism. So, it seems, at least two-thirds of respondents were not disposed to retract ascriptions of freedom upon learning the truth of determinism.

At least two-thirds; it may be more. We cannot conclude that those who responded that the agent did not act freely *are* disposed to retract. They might believe (i) that we have free will, (ii) that free will is incompatible with determinism, and (iii) that determinism is false — and they might be disposed, upon coming to learn that determinism is true, to reject (ii) and hold on to (i).[13] (Van Inwagen [1983: 219–220] has said explicitly that he endorses (i)–(iii), and would keep (i) and reject (ii) upon learning the falsity of (iii)). People with this cognitive makeup do not have the relevant dispositions to retract, even though they can be expected to respond 'not free' to the vignettes.

We have considerable evidence, then, that people do not have the relevant dispositions. The evidence is not *conclusive*; perhaps further empirical work will strengthen the case that we in fact have these dispositions.[14] But the preliminary evidence looks strong, and gives the friends of the Ideal Interpreter Argument hope.[15]

References

Daw, Russell and Torin Alter 2001. Free Acts and Robot Cats. *Philosophical Studies* 102: 345–357.

Devitt, Michael 1981. *Designation*. New York: Columbia University Press.

——— 1991. Naturalistic Representation. *The British Journal for the Philosophy of Science* 42(3): 425–443.

Fischer, John Martin and Mark Ravizza 1998. *Responsibility and Control*. Cambridge: Cambridge University Press.

an agent's birth, everything about their subsequent life; in another, respondents are told that the "beliefs and values of every person are caused completely by the combination of one's genes and one's environment" [Nahmias *et al.* 2005: 573].

[13] Thanks here to Ross Cameron.

[14] Shaun Nichols and Joshua Knobe [2007] critique Nahmias et al.'s conclusions (and are responded to in Nahmias *et al.* [2007]). One relevant issue: Nichols and Knobe's probes asked *counterfactual* questions — questions about 'universes' where determinism is true — whereas Nahmias *et al.* asked respondents to 'suppose scientists discovered' the truth of determinism. The latter are more directly relevant to dispositions to retract.

[15] Thanks to Ross Cameron, Cynthia Camp, Mark Heller, Eddy Nahmias, Helen Steward, Robbie Williams, an anonymous referee, and audiences at Stanford University and the Universities of Leeds and Regensburg. Research supported in part by the British Academy, grant SG-53931.

Flew, Antony 1955. Divine Omnipotence and Human Freedom. In *New Essays in Philosophical Theology*, eds. Antony Flew and Alasdair McIntyre, London: SCM Press.

Frankfurt, Harry 1971. Freedom of the Will and the Concept of a Person. *The Journal of Philosophy* 68(1): 5–20.

Heller, Mark 1996. The Mad Scientist Meets the Robot Cats: Compatibilism, Kinds, and Counterexamples. *Philosophy and Phenomenological Research* 56(2): 333–337.

Kane, Robert 1996. *The Significance of Free Will*. New York: Oxford University Press.

Kripke, Saul 1972. *Naming and Necessity*. Harvard University Press.

Lewis, David 1974. Radical Interpretation. *Synthese* 23: 331–344. Reprinted, with postscripts, in Lewis [1983b]: 108–118.

——— 1983a. New Work for a Theory of Universals. *The Australasian Journal of Philosophy* 61: 343–377. Reprinted in Lewis [1999]: 8–55.

——— 1983b. *Philosophical Papers*, vol. 1. Oxford: Oxford University Press.

——— 1984. Putnam's Paradox. *The Australasian Journal of Philosophy* 62: 221–236. Reprinted in Lewis [1999]: 56–60.

——— 1999. *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.

McKenna, Michael 2008. A Hard-Line Reply to Pereboom's Four-Case Argument. *Philosophy and Phenomenological Research* 77(1): 142–159.

Nahmias, Eddy, D. Justin Coates, and Trevor Kvaran 2007. Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions. *Midwest Studies in Philosophy* 31: 214–242.

Nahmias, Eddy, Stephen Morris, Thomas Nadelhoffer, and Jason Turner 2005. Surveying Free Will: Folk Intuitions about Free Will and Moral Responsibility. *Philosophical Psychology* 18(5): 561–584.

——— 2006. Is Incompatibilism Intuitive? *Philosophy and Phenomenological Research* 21(5): 597–609.

Nichols, Shaun and Joshua Knobe 2007. Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions. *Nous* 41: 663–685.

Pereboom, Derk 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.

Putnam, Hilary 1962. It Ain't Necessarily So. *The Journal of Philosophy* 59: 658–671.

——— 1975. The Meaning of Meaning. In *Mind, Language, and Reality*, vol. 2, Cambridge: Cambridge University Press.

Sterelny, Kim 1983. Natural Kind Terms. *Pacific Philosophical Quarterly* 64: 110–125.

Strawson, P. F. 1962. Freedom and Resentment. *Proceedings of the British Academy* 48. Reprinted in Strawson 1974.

——— 1974. *Freedom and Resentment and Other Essays*. Methuen.

van Inwagen, Peter 1983. *An Essay on Free Will*. Oxford: Oxford University Press.

——— 1989. When Is the Will Free? *Philosophical Perspectives* 3: 399–422.

Weatherson, Brian 2003. What Good are Counterexamples? *Philosophical Studies* 115: 1–31.

Yablo, Stephen 2002. Coulda, Woulda, Shoulda. In *Conceivability and Possibility*, eds. Tamar Szabò Gendler and John Hawthorne, Oxford: Oxford University Press, 441–492.