Acta Analytica (2020) 35:469–492
https://doi.org/10.1007/s12136-019-00415-y

Is Metaphysics Immune to Moral Refutation?

Alex Barber[1]

Received: 3 July 2019 / Accepted: 12 November 2019 / Published online: 5 December 2019

Abstract
When a novel scientific theory conflicts with otherwise plausible moral assumptions, we do not treat that as evidence against the theory. We may scrutinize the empirical data more keenly and take extra care over its interpretation, but science is in some core sense immune to moral refutation. Can the same be said of philosophical theories (or the non-ethical, 'metaphysical' ones at least)? If a position in the philosophy of mind, for example, is discovered to have eye-widening moral import, does that count against it at all? Actual responses by philosophers to the question of whether unanticipated moral consequences of metaphysical theories have evidential force are scattered, implicit, divergent, under-argued, and sometimes even self-undermining. The present discussion is, most immediately, an attempt to sort out the confusion. Beyond that, it exploits the new perspective this question gives us on a familiar topic: the relation of philosophy to science.

Keywords Naturalism · Metaphysics · Metametaphysics · Moral immunity

1 Introduction

To what extent and in what respects should philosophy assimilate itself to science? In this discussion, I aim to shed new light on this familiar question by addressing a less familiar one: to what extent if any is philosophy immune to moral refutation? Science *is*, it seems, immune to moral refutation. Moral opinion, no matter how profoundly held, has no evidential force in the empirical domain. As I argue elsewhere, even moral realists are unwilling to use the moral import of a scientific theory as evidence for or against it. Is philosophy also immune to moral refutation? The question has only been addressed in a piecemeal way, always in the context of dealing with the unexpected ethical implications of some particular philosophical position. The negative claim in what follows is that these isolated discussions are unsatisfactory, mainly because they

✉ Alex Barber
alex.barber@open.ac.uk

[1] Philosophy Department, The Open University, Milton Keynes MK7 6AA, UK

fail to consider the question at a general level. My first positive goal, then, is to tackle the question head-on. My second is to use the topic of immunity to moral refutation to develop a fresh perspective on the subtle relation between philosophy and science.[1]

As per the title, I will talk of 'metaphysics' rather than 'philosophy'. Moral philosophy is susceptible to moral refutation more or less trivially, so I will set it aside by using the word 'metaphysics', stipulatively, to describe all those parts of philosophy not generally seen as sub-branches of ethics. Metaphysical theories in this (loose but functional) sense can have moral ramifications, certainly, or the question of their moral refutability would never arise; but as with scientific theories, this moral force is not overt within the theory.

Explicit discussion of immunity to moral refutation (henceforth: moral immunity) is rare in the philosophical naturalism and metaethics literatures. There is plenty on whether science can inform metaphysics, and on whether metaphysics can inform ethics. Science can also have direct ethical consequences, unmediated by metaphysics.[2] Harder to find is any systematic consideration of inferences flowing in the other direction—from ethics to metaphysics and science.[3] Yet if science can inform metaphysics and both can inform ethics, what if anything stands in the way of ethics informing the other two, by the application of *modus tollens*? Raising this question has the potential to tell us a good deal about the structure of justification across the trio of science, metaphysics, and ethics.

To illustrate how this question arises, I begin with two case studies (Sections 2 and 3). Other examples are given throughout the paper, but these first two reveal the inconsistent ways authors react when a metaphysical theory is found to have decidedly odd ethical implications. My strategy thereafter is to draw on a pre-existing explanation of *science's* immunity to moral refutation (sketched in Section 4 but taken from Barber 2013), then ask whether that explanation carries over from science to metaphysics (Section 5 onwards). No simple answer—'Yes it does' or 'No it does not'—holds water, but the investigation nonetheless pulls us towards a version of naturalism that gives a credible account of the evidential relations between science, metaphysics, and ethics.

2 First Case: Would It Be Okay to Eat Swampman?

For our first example of a metaphysical theory (in the sense defined) with unexpected ethical consequences, consider teleosemantic theories in the philosophy of mind. These treat the content of a mental state in terms of the state's function, in a sense of 'function' that is ultimately naturalistic and non-purposive. Teleosemantic theories are sometimes taken to fall foul of the 'Swampman' objection. Swampman, physically identical to an actual person, is produced by accident in a lightning strike that coincidentally kills the original (Davidson 1987, pp. 443–4). Intuitively, this duplicate has the same contentful states as ordinary humans,

[1] I discuss science's moral immunity in Barber (2013). For some crucial caveats, see note 11 below.

[2] For science-to-metaphysics, see, e.g., Hawley (2006), Chalmers et al. (2009), Ross et al. (2013); for metaphysics-to-ethics, see, e.g., Stern (1992) (plus references therein to an earlier disagreement between Derek Parfit and John Rawls), Conee (1999), Carpenter (2014). For science-to-ethics: even Cohen (2003), who argues for a category of 'fact-insensitive' ethical principles, allows that empirical knowledge can be decisive in the ethical domain.

[3] The Harman/Sturgeon debate (e.g. Harman 1977; Sturgeon 1988) is only an apparent exception. The real interest there is in the legitimacy of inferences *from* science *to* ethics, not vice versa. It sometimes seems to run in the other direction because the inference-type at issue is inference to the best explanation, meaning the debate turns on whether ethical claims can explain non-ethical phenomena.

but teleosemantic accounts cannot accommodate this intuition. Swampman lacks the history (evolutionary or otherwise) needed to ground naturalistic functions.

Teleosemanticists have typically responded by dismissing or overriding the intuitions we might have about the scenario. David Papineau adopts this approach in an early discussion. He says he is offering a scientific reduction, not analysing a folk concept (Papineau 1993; see also Millikan 1996). But later he recounts an interesting twist that brings us to our theme. A student asked him whether it would be okay to kill and eat Swampman for lunch. Papineau (2001) writes that the objection 'stopped me in my tracks' because 'when we are forced to consider the ethical consequences of this decision, then we seem to end up with the wrong answer. If we did come across a Swampman, it would clearly be wrong to kill it for meat' (p. 282). Overriding this ethical intuition, as he and others had earlier overridden non-ethical intuitions, was not an acceptable option. At first, he thought he could get around the problem on the grounds that Swampman's non-intentional states (its susceptibility to pain and suffering, say) would render it inedible; but that would put it on the same moral footing as a cow, and eating Swampman, it seemed to Papineau, would be worse than eating a cow.

Papineau eventually settles on a teleolosemanticist response to this new, moral version of the Swampman objection. Instead of disregarding the recalcitrant intuition, as he had done with non-moral ones, he finds a way of reconciling it with a qualified version of his theory. He continues to claim that his theory is a 'substantial scientific claim' (2001, p. 286), but rather than using this scientific status as a reason to dismiss the ethical intuition, he says that his theory is a claim about the *actual* realizers of functional roles, not about alternative possible realizers in non-actual circumstances (e.g. Swampman in the thought experiment).

The point I want to stress here is that Papineau takes the moral threat to his position seriously. He acknowledges the challenge generated for his theory by its apparent inconsistency with a plausible moral claim. This acknowledgement seems to be in tension with his claim to be offering a scientific theory. He could have just said that science trumps ethics. No one would challenge a geological or biological theory *simply* on the grounds that it has questionable moral ramifications. To put it more strongly, we would treat such an 'objection' (2001, p. 282) as entirely wrong-headed. Empirical evidence, no matter how slim, beats moral evidence, no matter how robust. Why, then, does Papineau suppose that the edibility worry represents a potential objection to his theory? Does his doing so not hint at a lack of conviction in his proffered solution? It must be unclear at some level, even to him, that he is offering us a genuinely scientific theory. Otherwise, his response might have been: 'Mine is a scientific theory, and so is immune to moral refutation.' This is what we would expect to hear from marine biologists if, in attempting to settle the question of whether fish have a capacity to suffer, someone supplemented the usual behavioural and neurological evidence with 'moral evidence' extrapolated from the premise that it is morally impermissible to harvest and eat fish.

3 Second Case: the Moral Considerability of Counterparts

Our second case study of a metaphysical theory with unexpected moral import is modal realism, the view that possible people, possible things, possible events, etc., are no less real than actual ones, even if they are spatiotemporally and causally inaccessible to us.

This has the counterintuitive implication that we are under no obligation to do, or to refrain from doing, anything whatsoever. Or rather, it implies this when it is combined with utilitarianism, but it has similarly troubling consequences when allied to other ethical frameworks. This result looks like an invitation to draw a non-moral conclusion—that modal realism is false—from moral premises. Should we accept such an invitation? After setting out the case for the entailment in more detail, I describe two very different ways of handling it.

David Lewis's canonical statement of the position can serve as our default (Lewis 1986a). The core of his view is that non-actual possible worlds are as real as the actual world but are causally, spatially, and temporally isolated from it and from each other. To this core, Lewis adds two further pertinent claims. First, the expression 'actual' is an indexical, picking out the world in which it is uttered. A person or a moment is not made more real than other people or other moments merely by being the referent of an utterance of 'I' or 'now'; likewise, the actual world is not made more real merely by being the possible world in which an utterance of 'actual' happens to be produced. Second, occupants of non-actual possible worlds can be counterparts of, but are never numerically identical with, occupants of the actual world.[4]

The troubling moral consequences for the view emerge most starkly against a utilitarian backdrop. From Jeremy Bentham onwards, utilitarians have rejected as arbitrary what others have seen as legitimate barriers to moral consideration. Discussing the treatment of non-human animals, Bentham (1789) famously stated that 'the question is not, Can they *reason*? Nor, Can they *talk*? but, Can they *suffer*?' (p. XVII.1n). We might paraphrase this with: 'The question is not, Are they *actual*? but, Are they *real*?' After all, if 'actual' is a mere indexical then the as-real-as-us inhabitants of other possible worlds seem entitled to join non-Europeans, non-males, and non-humans in utilitarianism's ever-expanding nation of morally considerable beings. Awarding moral status to non-actuals, however, results in *moral paralysis*, the view that we have no moral obligations one way or the other. Any apparently utility-promoting act I perform is matched by my counterpart's non-performance of a type-identical act in a world that is otherwise similar; equally, my non-performance of it would be matched by my counterpart's performance of it. Whatever my choice, the act will (really) be performed once and once only, whether by me or by my counterpart, and the net change in real utility will be zero.

There is a tempting reply to this reductio argument. Suppose we represent the argument as follows:

> P1 Nothing we do will alter the sum of utility across both actual and non-actual worlds (background assumption[5]).
>
> P2 Non-actual and actual utility are equally real (from modal realism).
>
> Sub-conc. Nothing we do will increase real utility.

[4] x is a counterpart of another world's y iff x is similar to y along contextually salient dimensions.
[5] Cf. Lewis (1986a, p. 126): 'the character of the totality of all the worlds is not a contingent matter'.

> P3 We are morally required to perform only those acts that maximally increase real utility (from utilitarianism).
>
> Conc. We are not morally required to do anything (moral paralysis).

What is to prevent us from adopting a version of utilitarianism—P3, but with 'actual' in place of 'real'—that would block the reductio?

We have seen one reason for thinking that utilitarians should resist that move. Utilitarianism is by default an inclusive doctrine, even if that calls for revisions to our moral outlook. Recognizing non-human utility, for example, forces uncomfortable changes to our moral opinions, but that, to a utilitarian, is hardly grounds for only counting human utility. Likewise, we might think, for this proposal to only count actual utility. To overcome this presumption of inclusivity and revisionism, utilitarian modal realists need an independent rationale for amending P3.

A possible independent rationale is the principle that ought-implies-can. Since we cannot aid counterparts—we have no causal access to their world—we have no obligation to aid them; and if we have no obligation to aid them, we need not include them in our moral calculations. Lewis (1986a) himself thinks the problem sketched above only arises for a 'pure' sort of utilitarianism on which we are not restricted to helping only 'those whom [we] are in a position to help' (p. 128). This appeal to ought-implies-can, I believe, rests on a faulty analogy with the principle's familiar use in intra-world scenarios. I have relegated that suspicion to a long note.[6] Here, I will simply point out that the ought-implies-can principle is readily accommodated within a more fully specified version of P3:

> P3' We are morally required to perform *all and* only those acts that *it is within our capacity to perform and that* maximally increase real utility.

Replacing P3 with P3' would not undermine the reductio argument because P3 is logically entailed by P3'.[7] I used the stripped-down version because it is all the reductio argument requires for validity.

[6] Consider A, confronted with a situation in which B and C will both die if she does nothing. A cannot save both so must choose and is torn. Then she sees that she cannot save C anyway, whereas she can, and so does, save B. By ought-implies-can we commend A for acting as she did. In an apparent analogue of this intra-world scenario, X cannot save both Y (a worldmate of X) and Z (Y's counterpart in another world). X is causally isolated from Z's world so cannot save him anyway; by analogical reasoning, we should commend X for opting to save Y. But the analogical reasoning is faulty, and when we fix the fault we get the opposite result. While the intra-world scenario has one agent, A, the trans-world scenario has two, call them X* (formerly X) and X_C (X*'s counterpart). While we can allow that X* cannot *cause* anything to happen in any world but her own, she acts in a way that *metaphysically requires* Z's demise through X_C's inaction (cf. Heller 2003, pp. 8–9). A better intra-world analogue of this would have A in a position to save B but not C, as before; but A can save B only by thereby preventing some other actual person, A', from saving C. And here, a utilitarian (and many non-utilitarians for that matter) would be neutral on which option A should choose. This is exactly the wrong result if we want to argue by analogy that X* should act on behalf of her worldmate Y.

[7] The entailment is rudimentary: $\forall x$ (Fx iff (Gx & Hx)) $\vDash \forall x$ (Fx only if Gx). 'Morally required' does not, despite appearances, have wide scope in P3 or P3' (it is sealed within 'F'), so the entailment is not trading on any suspect features of deontic reasoning (e.g. Forrester 1984). Anyone nonetheless concerned about deontic paradoxes is referred instead to the response in the previous footnote.

Switching to non-utilitarian moral frameworks does not help here. Kant, for all that he was no utilitarian, was not opposed to means-end reasoning or to the promotion of welfare. Indeed, contributing to the happiness of others is Kant's (2002 [1785]) chosen example of an imperfect duty to others (p. 40). His view is just that means-end justifications are subject to, or conditional on, the more familiar apparatus of his moral philosophy, not that they are always mistaken. Kantian modal realists would thus face the same difficulty. Virtue ethics would likewise be compromised. Virtues like kindness, charity, and justice look highly suspect if the inevitable cost of my exercising any of them is that my counterpart exercises egoism, selfishness, or injustice. Any version of virtue ethics that trumpets personal integrity ('clean hands') over all else would be excessively smug, a vice in itself.

If we therefore accept that modal realism has difficult moral implications, what follows? I will now describe two diametrically opposed reactions from two authors, both of whom accept that modal realism implies what I am calling moral paralysis. Mark Heller thinks it shows we must reject modal realism. Torbjörn Tännsjö thinks it poses no such threat. Neither author argues effectively for their preferred stance.

Heller thinks it is a consequence of modal realism that we are under no obligation to save an easily saved drowning child. Here is how he argues that we should therefore reject modal realism

> If modal realism commits its adherents to behaving contrary to a moral truth, then it is to that extent immoral. And to the extent that the moral truth is obviously true, that gives us reason to believe in the falsity of any theory that conflicts with it. The objection to modal realism is therefore, not just that it is immoral, but that it is therefore false. (Heller 2003, p. 3)

The force of this objection is not as strong as Heller appears to think. Of the two charges against modal realism, i.e. immorality and falsity, the first is a distraction. It overcomplicates what is in fact a simple argument: modal realism is false because it implies a moral falsehood (e.g. moral paralysis, or else Heller's more specific example of its being okay to ignore an easily saved child). Once our attention comes to rest on this simple argument, however, his conclusion looks too quick; or rather, it rests too heavily on a proviso that is never made good ('*to the extent that* the moral truth is obviously true'). Why should we not, instead, conclude that we need to radically revise our moral outlook, by embracing moral paralysis or by abandoning some other part of our moral framework?

That, recall, would be our reaction in a scientific context. Consider modal realism's scientific cousin, the multiverse interpretation of quantum mechanics (Wallace 2012). Setting aside differences between it and modal realism, notice how peculiar it would be to try to settle the debate between competing interpretations by invoking moral considerations. These would not even constitute a tie-breaker. Heller assumes, in effect, that things are otherwise for metaphysical theories than they are for scientific ones. So, apparently, does Lewis himself, in so far as he takes the time to address the 'moral' challenge to his view in a way that goes well beyond simply noting that metaphysics trumps ethics. Lewis's and Heller's attitudes are not as puzzling as Papineau's, since they do not categorize modal realism as a scientific theory. But they still fail to consider the possibility that, *like* a scientific theory, it is immune to moral refutation.

Tännsjö, too, thinks modal realism is massively at odds with existing moral opinion; but unlike Heller, he recommends recalibrating our ethical norms rather than our modal metaphysics. Specifically, he claims that utilitarian modal realists can simply endorse what I am calling moral paralysis.

> Together with modal realism…, [utilitarianism] yields the conclusion that, morally speaking, anything goes [i.e. moral paralysis]. … This conclusion is unexpected of course, but, as such, it *does not constitute any good reason why we should give up modal realism* (or utilitarianism). Utilitarianism as such is at variance with several common-sense moral principles, but the conclusion that anything goes is not as such at variance with any plausible moral principle. We should not be prejudiced as to the exact scope of our moral principles. Perhaps there are very many actions that we want to perform that are morally prohibited, perhaps there are very few. The case where none is prohibited is a limiting case, no more and no less. … If, *pace* David Lewis, we believe that we have good reasons to accept pure utilitarianism, we may stick to it also when it turns out to imply that we have no moral obligations. This implication makes our reasons for accepting pure utilitarianism neither better nor worse than they were before we knew about it, *nor should we allow it to upset our belief, if we happen to entertain such a belief, in modal realism*. (Tännsjö 1987: 88-9, emphasis added)

The salient feature of this passage for our purposes is its lack of an argument for the claim that modal realism is safe, despite being inconsistent with the conjunction of utilitarianism and the denial of moral paralysis. Tännsjö asserts this twice in the passage (italics) but his reasons have to do only with whether we should embrace moral paralysis before we abandon utilitarianism. Whereas Heller, like Lewis, assumes that modal realism is susceptible to moral refutation in a way that science is not, Tännsjö assumes the opposite. There is a paucity of argument on either side.

4 Why Is Science Immune to Moral Refutation?

To settle the question raised by our two case studies and others like them, a sensible place to start is with a sense of why *science* is immune to moral refutation; we can then ask whether the same explanation extends to metaphysics, given parallels and discrepancies between it and science. I will therefore draw on my earlier (2013) explanation of science's immunity to moral refutation. After describing that explanation in this section, I will address its extendibility to metaphysics in Section 5 onwards.[8]

Let us return to the fish example. Suppose a marine biologist maintains on empirical grounds (e.g. behavioural and neurological traits) that fish have the capacity to suffer. Now imagine an attempt to refute this position using moral premises and a valid-looking[9] inference:

[8] In the earlier paper, I was concerned with the question of whether science's immunity to moral refutation means we should be moral anti-realists. I will sidestep that metaethical question here by accepting moral realism and simply taking over what I there argue is the only explanation of science's moral immunity a moral realist can accept.

[9] I have suppressed the deontic premises needed to make it formally valid, which I take to be harmless here.

> To harvest and consume creatures with the capacity to suffer is to generate unnecessary suffering; one ought not to generate unnecessary suffering; it is morally permissible to harvest and consume fish; therefore, fish lack the capacity to suffer.

This ethical contribution to the scientific enquiry will seem wrong-headed even to those predisposed to accept its three premises. They do not even yield a tie-breaker. What explains this?

For a moral realist—here understood to be someone who accepts that moral discourse is a reasonably trustworthy source of true and objective judgements—this is, on the face of it, puzzling. Figure 1 is a visual representation of this puzzle. Given some logical inconsistency between a set of ethical claims (e.g. the moral premises in the 'fish' argument above) and some scientific claim (e.g. the view that fish have the capacity to suffer), why does evidence for the ethical claims (coming in from the right) never act as a counterweight to the empirical evidence for the scientific claims with which the ethical claims conflict (evidence coming in from the left)? In other words, why can moral evidence not be bundled up with more traditional kinds of evidence in the evaluation of science?

Figure 2 is a representation of my solution to this puzzle (2013, pp. 644–7). It involves accepting a sharp division within the set of ethical judgements between those the evidence for which is purely a priori and those that are derived from a conjunction of this first category plus our empirical assumptions. Crucially, the 'applied' ethical judgements in this second category have no evidence in their favour beyond their derivability.

Once we accept this division of justified ethical judgements according to the source of their justification, we can see that neither source can be mustered into any kind of counterweight to the empirical evidence for scientific judgements. Non-derived ethical judgements in isolation, with their purely a priori support, can never conflict with empirical findings. After all, if they had this kind of empirical significance, they would themselves be open in principle to empirical assessment, meaning the evidence for them was not purely a priori after all. As for derived ethical judgements, they have only derivational support, and the source of that derivational support is, in part, empirical. To draw empirical conclusions from them would therefore always involve circular

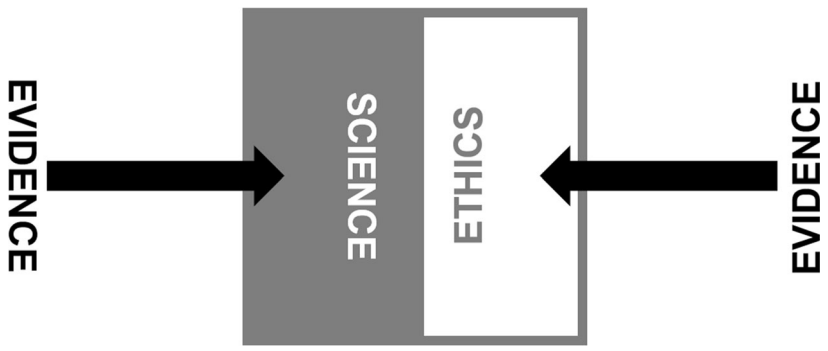

Fig. 1 A puzzle: why does the ethical evidence have no force against the scientific evidence?

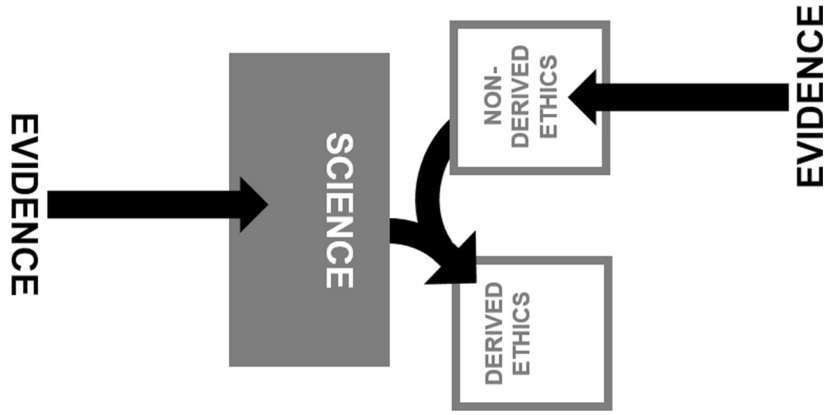

Fig. 2 A solution to the puzzle: science is immune to moral refutation because there is a division within the category of ethical judgements

reasoning. The fish argument above illustrates this circularity. The derived ethical judgement that it is morally permissible to harvest and consume fish (the third premise) is warranted *at most* to the extent that we already have grounds for thinking that fish lack the capacity to suffer.[10]

5 From Science's Immunity to Metaphysics' Immunity: Three Possibilities

Our next task is to search for parallels or discrepancies when we turn from science to metaphysics. This section is designed to narrow us down to three possible positions, to be selected from in the rest of the paper. Before any of this, I will set out four working assumptions. Any of these assumptions could be challenged but adopting them will keep a lid on a discussion that can become very complex very quickly.

Working assumption 1: *Science is immune to moral refutation*.[11]

[10] For simplicity I have ignored non-welfare arguments, e.g. environmental considerations. These too would rely on empirical assumptions, complicating the point without undermining it.

[11] Some caveats are in order. (i) Ethical considerations can legitimately shape a scientific *agenda* (e.g. cancer research). (ii) Sometimes we may wish to raise the evidence threshold for *acceptance* of a scientific theory because of its strong moral import (see, e.g., Kitcher 1985, p. 9). (iii) Sometimes the *ideological ramifications* of a finding can legitimately prompt us to suspect that 'decision-based evidence-making' rather than science is at work (see, e.g., Newby and Newby 1995). (iv) Sometimes we draw on *moral psychology* for evidence (see, e.g., Rini 2013), but this is not the same as using a moral premise as such to draw an empirical conclusion. In van Fraassen's (2002) words, 'once we are given the fact that the humans involved had certain values and made certain value judgments[, …] it is clear that not values but only humans' having values is what really matters' in science (p. 182).

Science's immunity is, moreover, strict: it is not just that moral evidence counts *only a little* against a scientific position.[12]

> Working assumption 2: *The explanation of science's immunity sketched in Section 4 is correct.*

Ethics, that is, contains an evidentially defined split between a purely a priori part and an entirely derived part.[13]

> Working assumption 3: *If metaphysics has moral immunity, this fact has the same explanation as science's moral immunity.*

The thought here is just that it would be too great a coincidence if, even though the explanation of science's moral immunity failed to carry over to metaphysics, some distinct generator of moral immunity for metaphysics was nevertheless waiting in the wings.

> Working assumption 4: *Moral realism is broadly correct.*

This is the most contentious assumption, and I make it only so as not to short-circuit the discussion. The moral immunity of both science and metaphysics is relatively banal if we grant moral anti-realism. Plenty of people are moral realists, however, and even those who are not can find interest in the questions moral immunity raises for those on the other side of the fence.

Let us turn now to whether metaphysics is immune to moral refutation for the same reason that science is. One might think that the explanation I have given of science's moral immunity can be extended quite easily: just slot in the phrase '…and metaphysics' after 'science', and exploit the same immunity-generating split within ethics between its purely a priori part and its derived part. Derived ethics will now include judgements that result from combining purely a priori ethical principles with scientific *and/or metaphysical* theories. On this proposal, just as the ethics of harvesting fish can tell us nothing about neuroscience or ethology, so Swampman-ethics can tell us nothing about the philosophy of mind, counterpart ethics can tell us nothing about modal metaphysics, and so on. To suppose otherwise would be to open oneself up to the same charge: circularity. Applied ethical claims (e.g. 'Eating Swamp-people is wrong') would have no epistemic support beyond their derivability from, among other things, these self-same metaphysical theories (e.g. teleosemanticism).

This simple opening case for the moral immunity of metaphysics instantly faces two objections. The first is that Heller, Lewis, and Papineau have all shown themselves drawn to an alternative view: that metaphysics, unlike science, *is* susceptible to moral

[12] The case against moral refutations of science seems to be empirical rather a priori. There is nothing logically or conceptually incoherent about scientific theories with a value-rich ontology, but such theories have been gradually and comprehensively phased out in exchange for greater empirical adequacy. As Wilson (1995) puts it, in early modern science 'the subtraction of moral attributes [led to a] new discourse of nature [that] was as impoverished in cultural terms as it was enriched in observational terms' (p. 212).

[13] Basl and Coons (2017) offer an interesting alternative to my explanation, but there is no space here to assess its implications for the immunity or otherwise of metaphysics.

refutation, at least in principle. They may be wrong in being so drawn—and Papineau, one suspects, would be happy to accept as much—but we would still have some obligation to explain what draws them at all. They do not react with ridicule, as anyone would to moral 'reasoning' about the nervous systems of fish. If both science and metaphysics are immune to moral refutation, then either the explanations of this immunity differ (contrary to working assumption 3) or else the explanations are the same but some yet-to-be-unveiled supplementary explanation is needed for the failure by Heller and company to spot this sameness.

The second objection is that placing metaphysics alongside science in this way sabotages the explanation of moral immunity sketched in Section 4 (contrary to working assumptions 2 and 3). A crucial posit of that explanation is a category of ethical claims that have purely a priori justification and hence cannot come into evidential conflict with empirical claims. If metaphysical knowledge is wholly or partially a priori, that part of the explanation will not necessarily carry over.

We can deal with both objections if we strengthen the proposal: instead of saying that metaphysics is immune to moral refutation for reasons *parallel* to those that make science immune, we can say it is immune because, done properly anyway, metaphysics *is* science. Or rather (and to avoid conjuring up unintended images of philosophers in protective goggles), because it is continuous with paradigmatic instances of science in certain vital respects, such as being beholden to the same standards of empirical adequacy, the same principles of explanatory coherence, the same ontology, or whatever it may be. This modification, this naturalism about metaphysics, yields responses to our two objections.

Regarding the first objection, hesitancy to dismiss moral objections to metaphysics can be interpreted as hesitancy over the naturalism such a dismissal presupposes. Either the hesitators are not philosophical naturalists at all, or they are but not dogmatically so. Anything short of knee-jerk certainty that metaphysics (done properly, etc.) is an extension of science will leave room for a suspicion that metaphysics is inferentially entwined with ethics in a way that exposes it, unlike science in the narrow sense, to morally driven revision. Such suspicions could give rise to searches for responses to ethical objections to metaphysics—redundant searches if naturalism is in fact correct.

Regarding the second objection, once we treat metaphysics as part of science, there is no need to worry about metaphysics having an a priori element. On a naturalistic conception of metaphysics, the explanation of science's moral immunity could be extended quite straightforwardly into an account of metaphysics' moral immunity. All evidence for applied ethical judgements is derivational, so using it against science—now in the wider sense, incorporating metaphysics—would be circular; and the purely a priori evidence for non-derived ethical judgements entails that they never conflict with science, again in this new wide sense.[14]

[14] A more complex naturalistic position than the one sketched here is also possible. I have been envisaging an especially empirically minded naturalist for whom the norms of scientific enquiry (coherence with other scientific theories, consilience, explanatory elegance, simplicity, etc.) are themselves in some sense empirical rather than a priori. The more complex position would involve drawing and insisting on a distinction, something like Kant's, between practical and theoretical (including mathematical) knowledge. Both kinds could have a priori components—meaning science's epistemic norms could be a priori—so long as there was no chance of a priori moral norms being cited in evidence when we reason about epistemic norms. (If they were so cited, ethical evidence in the scientific domain would become possible, contrary to Working assumption 1.) In this exploratory paper, I will not try to settle whether science itself has an a priori element, ringfenced somehow from practical a priori knowledge.

This gives us an early tentative result in our enquiry into the moral immunity or otherwise of metaphysics. Two theses—naturalism about metaphysics and the thesis that metaphysics has moral immunity—appear to be interdependent. If we accept or reject one, it seems, we must do the same with the other.

That naturalism and immunity are biconditionally tethered in this way is not so surprising an outcome once it is made explicit. The biconditional leaves us with two positions to choose between, however: endorse both a naturalistic attitude towards metaphysics and the moral immunity that comes along with this naturalism; or reject both. And here we face an apparent impasse: the only way of deciding between these two options is to draw on our existing sympathies or antipathies towards naturalism. In that case, our choices will reflect those sympathies and antipathies but will not help us to evaluate them. This threatens to expose as mere hype a claim made at the start of this paper: that reflecting on moral immunity opens up a valuable new perspective on the naturalism debate.

We can get beyond this impasse, though our situation will get worse before getting better. I am about to argue that *neither* of these two options is acceptable; but out of their ashes, a third option will emerge, a pluralistic form of naturalism. Call the first two options *across-the-board naturalism* and *across-the-board anti-naturalism* respectively. Across-the-board naturalism holds that metaphysics is *all* naturalistic and so *all* immune to metaphysics; across-the-board anti-naturalism holds that metaphysics is *all* non-naturalistic and so *all* susceptible to moral refutation. I argue against these views in Sections 6 and 7. On the pluralistic approach I prefer, we do not have to be naturalistic about all of metaphysics in order to be naturalistic about some. I elaborate this third option in Section 8, defending it against an important objection. I may appear to be attacking figures of straw in bothering to reject the other options first, but discussions of naturalism do tend to be conducted as if one has to be either pro-naturalist across the board or else anti-naturalist across the board.[15] It is therefore worth showing explicitly that a pluralist stance is not merely an option, it is in fact required.

6 The Case Against Across-the-Board Anti-naturalism

The loose statement of naturalism I gave earlier—that metaphysics 'is continuous with paradigmatic instances of science in certain vital respects'—glosses over familiar divisions within the naturalist camp. These divisions are about to become important. Some variants of naturalism are more credible than others. To give two extreme (and made-up) examples: a variant that calls for the dismissal of all metaphysical theses that have not been semantically reduced to the principles of current physics would have very low incredibility, whereas a version that commands philosophers to consult scientists when their own work has non-trivial empirical commitments would have very high credibility.

I will exploit this variability in credibility to argue against across-the-board anti-naturalism. (This, to repeat, is the view that we should take a non-naturalistic stance towards all metaphysical theories, and so view all metaphysical theories as susceptible

[15] Papineau's (2016) review, for example, helpfully distinguishes many subtypes of naturalistic attitude, but does not consider whether any one of them could be apt in some metaphysical debates but not in others.

Table 1 Four possible stances towards the evidence available in metaphysical debate

	Metaphysical theses are ripe for…	
	…empirical justification or refutation	…a priori justification or refutation
Stance A	✗	✗
Stance B	✓	✗
Stance C	✓	✓
Stance D	✗	✓

to moral refutation.) Adopting across-the-board anti-naturalism, I will show, means rejecting even a relatively plausible version of naturalism. I will call this plausible version of naturalism *weak* naturalism to signal how undemanding it is. First, I will explain what weak naturalism is and why it is appropriate to call it undemanding. Then I will show why weak naturalism is ruled out by across-the-board anti-naturalism

There are various conceptions one might have of the evidential relationship between metaphysical debates and science. Weak naturalism is a disjunction of such conceptions. It is weak in part because it is a disjunction, but also because it makes no overt ontological demands (reductionist or otherwise). To see what the disjuncts are, suppose we make a four-way division of stances one might take on the evidence-base for metaphysics (A to D in Table 1). One might hold that we should just do science in the narrow sense, abandoning traditional metaphysical debates as beyond meaningful evidential support of any kind (stance A). One might hold that enquiry into metaphysical topics is legitimate but only if it can be incorporated into science and dealt with empirically (stance B). One might hold that metaphysics can be informed, potentially anyway, by both empirical and a priori considerations, with trade-offs between these two sources (stance C). Finally, one might think that metaphysics is a legitimate domain of purely a priori enquiry, disengaged in principle from empirical considerations (stance D). By 'weak naturalism' I mean the disjunction of A, B, and C—in other words, the rejection of D.[16]

Weak naturalism about metaphysics is a relatively undemanding brand (or disjunction of brands) of methodological naturalism. Each disjunct within it, moreover, requires metaphysics to be immune to moral refutation—or so I am about to argue in the remainder of the present section. Weak naturalism therefore requires this too, by disjunctive syllogism. That represents a problem for across-the-board anti-naturalism: maintaining that metaphysics is always susceptible to moral refutation means rejecting even this relatively undemanding naturalism, across the board.

The incompatibility of both A and B with moral refutations of metaphysics can quickly be shown using our working assumptions. Suppose for reductio that metaphysics is susceptible to moral refutation. Working assumptions 1 and 2 commit us to a

[16] The purpose here is to divide the logical space for the disjunctive inference that follows, not to categorize actual authors. Quine, for example, adopts A and B at different times. That said, for representative expressions indicating sympathy for each of the alternatives, see van Fraassen (2002) (option A); Ladyman and Ross (2007) (option B); Lowe (2006), Paul (2012) (option C); Fine (2012) (option D).

split between non-derived and derived ethics, neither of which can be a source of evidence against science. But it is hard to see how either could be a source of evidence against metaphysics either, if A or B is correct. Non-derived ethics is a priori, so using *it* to refute metaphysics would be contrary to both A and B; derived ethics, on the other hand, because it is derived, is not an independent source of evidence.

C's empiricism is more forgiving than A's and B's, so C is less readily shown to be incompatible with moral refutations of metaphysics. The incompatibility can be shown nonetheless, again using our working assumptions. The short version is that if C is accepted, any moral refutation of metaphysics would carry over—by hypothetical syllogism—into a moral refutation of the scientific underpinnings of this metaphysics, in violation of working assumption 1. To show that this charge sticks, I will run through it in the abstract, then illustrate it with an example.

Suppose a metaphysical theory is susceptible to moral refutation. This refutation would have to be a priori in nature (for the reasons just set out in the discussion of A and B). That in itself poses no threat to C because C, unlike A and B, permits a priori evidence for metaphysical claims. The difficulty for C arises because the refutation would be ethical in nature, not that it would have an a priori source. This ethical nature is a problem because, on stance C, metaphysics is evidentially intertwined with science, so if metaphysics is susceptible to refutation on ethical grounds, so too is science—contrary to working assumption 1.[17]

To illustrate the problem generated for C if we permit moral refutations of metaphysics, consider a topic sitting within the purview of both physics and traditional metaphysics: time. Presentism, the view that the past and future are unreal, has sometimes been criticized as incompatible with relativity theory's abandonment of simultaneity (see Balashov and Janssen 2003, cf. Hawley 2006). But alongside this alleged conflict with an empirically supported scientific theory, presentism has—or has been held to have—challenging moral implications. Delmas Lewis, for example, claims that it is incompatible with holding individuals responsible for earlier actions and should be rejected for that reason.[18] We can imagine stringing together these two dialectic aspects of presentism—its empirical commitments in one direction and its ethical ones in the other—to give, absurdly, a moral argument for relativity theory: individuals are morally responsible for their actions; moral responsibility implies realism about the past and so the rejection of presentism; rejection of presentism supports relativity. There are, of course, plenty of potential vulnerabilities in this reasoning chain. But our reaction to the argument should not be, 'There's a debate to be had here!' Rather, it should be that something has gone seriously awry. We reach the absurdity only if we allow that the metaphysical debate is evidentially intertwined with science in the way that C permits *and* that metaphysics is vulnerable to moral refutation.

Summing up, across-the-board anti-naturalism requires us to reject even a relatively mild, disjunctive version of methodological naturalism. It forces us towards stance D,

[17] One might try to block this argument by allowing that science can inform metaphysics while denying the converse. But this inferential asymmetry, while it would save the compatibility of stance C with moral refutations of metaphysics, is devoid of any rationale. It would be warranted only if metaphysical claims lacked independent, non-empirical support. While that is true on stance B, for example, stance C recognizes the existence of a priori support for metaphysics, meaning the asymmetry would be unprincipled.

[18] Lewis (1986b). Oaklander (1988) also takes the moral objection seriously enough to argue against.

the view (which I am here assuming is relatively unattractive) that metaphysical theories, of time say, must in principle shrink into the shadows of scientific irrelevance.

7 The Case Against Across-the-Board Naturalism

The other 'across-the-board' option holds that we should always go for both naturalism and therefore moral immunity when doing metaphysics. This view faces a different kind of problem: metaphysics *is*, sometimes anyway, uncontroversially susceptible to moral refutation.

A classic illustration of this is Locke's memory-based criterion for persistence of self. Locke is perfectly explicit that he is offering a 'forensic' account of personal identity. ('Forensic' in its strict sense means pertaining to law, but here we can read it more loosely to mean pertaining to ethics.) His purpose, he says, is to understand how punishment can be directed towards 'the same that committed those actions, and deserve that punishment for them' (Locke 1997 [1706], p. 312). It follows that if an absurd assignation of moral responsibility emerges from his memory criterion of persistence, his memory criterion is in trouble. Suppose, for example, that a re-formed criminal describes himself as still having vivid memories as of committing a hideous crime, and at the same time sincerely and comprehensively disowns 'the person I once was' (as he puts it). We may feel we can no longer hold him responsible. In that case, we would need to give up on the memory criterion as a measure of personal identity in Locke's forensic sense (irrespective of whether we think there is persistence of self in some non-forensic sense). Locke could hardly shrug the objection off by pleading moral immunity when moral significance is his enquiry's stated *raison d'être*.

It is tempting to dismiss this objection to across-the-board naturalism as showing only that Locke's theory is not a metaphysical theory after all, but rather a barely disguised ethical theory. Locke's theory is not the sole example, however, and other examples show—in a way Locke's does not—that whether a metaphysical theory is purely science-oriented (so that any ethical implications it may have cannot be held against it) or ethics-oriented (and so accountable to the moral sphere) is not always easy to spot. Such cases are therefore less easily dismissed.

Before describing two philosophical cases, it will be instructive to consider a non-philosophical analogue. Sound level is not the same as noise level. Sound level is the pressure of an acoustic wave, standardly given in decibel units relative to a base level of 2×10^{-4} microbars. This otherwise arbitrary choice of base level reflects human interests: it approximates the average lower bound of human audibility. But setting aside how it is expressed, sound level is independent of human interests and is measurable as such. Indeed, acoustic science is integral to branches of modern engineering unrelated to human perception, such as the structural integrity of buildings. Noise level, on the other hand, is shot through with evaluative considerations that render it unfit for use outside the specific contexts in which those considerations are salient, making it unmeasurable without reference to these same considerations. A train's distant hoot at night may be tolerable or even soothing, and so less noisy than a neighbour's irritating radio, despite sound level being the same in each case. A blackbird's singing can delight us until we realize that it is mimicking a car

alarm. The notion of sound level, we might say, is science-oriented whereas the notion of noise is ethics-oriented. If someone is irritated by a 'noisy' conversation, it can make sense to respond that it is 'not noise, it is just people conversing'. The equivalent response to a measurement of sound level would make no sense.

This sound/noise distinction is easy to miss until it is pointed out. In metaphysics, too, it may not be obvious whether we are trading in noise-like or sound-like concepts. The philosophies of mind and language have both been seen by a majority of practitioners as essentially naturalistic endeavours, properly trading only in science-oriented, sound-like concepts. And yet examples can be found of ethical drivers that sit uneasily with this orthodoxy.

An example in the philosophy of mind is Andy Clark and David Chalmers' suggestion that mental representations can sit outside our craniums. The evidential pros and cons they offer for the extended-mind thesis do not include any moral ones. Others have spotted, however, that their thesis makes some instances of data-theft equivalent to kidnapping, or that deliberately wiping a person's digital devices would be quasi-murderous rather than simply damaging of property. Both considerations raise the possibility of taking a forensic approach to tracing the mind's borders. On this approach, we would aim to include whatever we think warrants protection through personal privacy or anti-assault legislation. Any specification of these borders would then be susceptible to ethical evaluation.[19]

In philosophy of language and linguistics, the topic of the semantics/pragmatics boundary—of what it is, of where it is, of whether it is—is usually approached with explanatory payoff as the accepted arbiter: in effect, the better the empirical fit, the better the account.[20] Jennifer Saul, however, has recently looked at the boundary through an ethical lens. Just as Locke wanted a theory of the same-person/different-person distinction that mapped onto the ethical contrast between being and not being culpable, Saul wants a version of the semantics/pragmatics boundary that maps onto the ethical contrast (or alleged contrast) between lying and merely misleading. Moreover, she argues, no existing account of the boundary is up to the task, so she offers a new one. In ecumenical spirit, Saul does not insist that these other accounts are misguided. They are just designed to different ends. In her opinion, then, this is less an instance of a hitherto ethical dimension suddenly being revealed as key to a debate, and more a conscious change of topic.[21] More recently, though, Robert Stainton has taken a bolder step. He claims that the missing dimension in discussions of the semantics/pragmatics boundary is *precisely* this ethical one. The distinction between an act of 'full-on stating' as opposed to mere insinuation—the semantics/pragmatics distinction in other words—comes down to a difference in moral force. As he puts it, 'full-on stating has a special *forensic* status' [my emphasis] in that it is a special-purpose device, one function of which is to make the speech act 'lie-prone' (Stainton 2016, pp. 405–6). If he is right, any discussion of the distinction will be susceptible to support or criticism on ethical grounds, just as Locke's theory of personal persistence is. But unlike Locke's theory, it

[19] Clark and Chalmers (1998). For ethical worries, see, e.g., Levy (2007), pp. 59–63.
[20] See, e.g., contributions to Szabo (2005) and Stojanovic (2008).
[21] Saul (2012), Chs. 2–3. Saul ultimately denies the existence of a morally significant lying/misleading distinction.

is something of a surprise that the notion of asserting should be noise-like in this way, rather than sound-like.

We can see now why dismissiveness towards the Locke counterexample to across-the-board naturalism (that it is 'not really metaphysics') will not do. While Locke's explicit use of the word 'forensic' could be used to argue that he is openly and deliberately doing ethics, in the other cases, the core notions were not foreseen to be forensic and may well turn out not to be so. This demonstrates that, for at least some philosophical concepts, we cannot always know whether it is best to treat them as noise-like or as sound-like and hence whether a naturalistic or forensic conception of their status is most appropriate. Responding to cases like the semantics/pragmatics boundary or the extended-mind thesis by insisting that they and any others that crop up are not really metaphysics would mean across-the-board naturalism has collapsed into the unhelpful thesis that philosophy is best approached naturalistically (and hence non-forensically) unless it is not best so approached. Of any particular case, we would not know whether to dismiss a moral challenge.

8 A Pluralist Alternative

Short of giving up one of the working assumptions, it seems we need an alternative to the two across-the-board options. Since these are the only two options compatible with the biconditional arrived at in Section 5 (roughly: naturalism about metaphysics if and only if moral immunity for metaphysics), we also need to revisit that. In this section, I will show how, by abandoning the letter but not the spirit of the biconditional, an attractive form of naturalism about metaphysics comes available. It makes good sense of the existence of what I will henceforth call forensic metaphysical theories (i.e. theories whose *raison d'etre*, whether or not this is recognized by their proponents, has to do with what they tell us about something in the ethical domain) but it does not require us to jettison naturalism across the board or to violate any of the working assumptions.

Let us start with the biconditional. We have already dropped the pretence that there is just one kind of naturalism. Now it is time to drop the pretence that one single approach should be taken to all metaphysical theories. Metaphysics does not have to be treated as an undifferentiated bloc, all or none of it an extension (in specified respects) of science, and all or none of it immune to moral refutation. Locke's theory and the other examples in Section 7 suggest that different kinds of metaphysics can co-exist. Not all ways of dividing up the world have to serve the singular ambitions of fundamental science, with categories tailored accordingly. When those ambitions are to the fore, we should indeed see the relevant theory as immune to moral refutation. But that fact is compatible with a new, more nuanced formulation of the biconditional: of any given metaphysical theory, *it* has moral immunity if and only if *it* is an extension of science.

In the rest of this section, I defend this new formulation of the biconditional, along with the pluralist view of metaphysics it permits, a view that treats some metaphysical theories as forensic (and so potentially susceptible to moral refutation) and others as non-forensic (and so immune to moral refutation). There is really only one barrier to accepting such a view, but it is a significant one.

The difficult emerges once we drop yet another pretence: that talking of a forensic metaphysical theory as ethics-oriented somehow implies that it is not also science-oriented. In reality, forensic theories face both ways. Locke's view rests on empirical assumptions about the integrity of memory and consciousness in ordinary circumstances. The extended-mind thesis grows out of functionalism, still the dominant explanatory framework in cognitive psychology. And as my brief summary indicated, even if the semantic/pragmatic distinction is in part an ethical project, contributions to that project will need to pay heed to findings in theoretical linguistics. Forensic metaphysical theories, then, have both ethical ambitions and empirical roots, with success overall requiring success on both fronts. This Janus-like character creates a problem. If a forensic metaphysical theory is undermined because of an apparent ethical inadequacy, why does this not translate (by hypothetical syllogism) into evidence against its scientific ground, i.e. into a moral refutation of science, in violation of working assumption 1?

To rid ourselves of this worry, we can set aside non-forensic metaphysical theories (including, presumably, most theories in the philosophies of physics, biology, etc.), since the worry only arises for forensic ones. For forensic metaphysical theories, however, we need a clearer model than we currently have of how they can be both ethics-facing and science-facing without thereby licencing moral refutations of science.

The alternative model I propose turns on an essential feature of non-derived (a priori) ethical claims: they are, or yield, conditional propositions that allow us to derive applied ethical claims from claims about the physical world. Without such conditionals, applied ethics would be groundless and a priori ethics would be pointless. The idea that normative ethics should supply conditionals with worldly propositions in their antecedents and practical imperatives in their consequents is a familiar one.[22] It is also implicit in my explanation of science's moral immunity (see Fig. 2 in Section 4): non-derived ethics is needed to licence any inference from scientifically credible claims (the antecedents of the conditionals) to derived ethical judgements (their consequents). Despite this familiarity, we would be hard pushed to specify 'if-then' statements with antecedents expressed in the language of some foundational branch of science and consequents stating clear practical directives. The sheer intractability of such a proposal is obvious. There have to be stepping stones along the way, breaking the journey. On what I will call the *stepping stone* model of forensic metaphysics, forensic metaphysics sits on this long conditional pathway, mediating between science and both parts of ethics.

The easiest way to conceive of this model is in terms of its applicability outside philosophy. Concepts such as *noisy*, *danger*, *table*, *drunk-driver*, *chaste*, or *risk* sit on this same pathway, providing us with 'stepping stones' midway between the physical sciences and the outputs of practical reasoning. Unless they are given artificially precise definitions, these concepts have unwieldy or loose physical application conditions; but they are better adapted for use in practical reasoning than they could possibly be if they

[22] Utilitarians, for example, offers a single conditional schema: if current circumstances are such that Ø-ing would promote utility, then one ought to Ø. Kant's maxims, which are what his universalizability test is meant to evaluate, take the form: if circumstances are such and such, then do this or that. Virtue ethics is less obviously productive of conditionals, but its promise rests on the thought that circumstance-sensitive decisions are well taken when they manifest possession of the virtues. So while the virtue approach itself will not always yield the kinds of conditionals I have in mind here, virtuous individuals operate with them.

were geared to the explanatory needs of science. Sitting alongside these non-philosophical concepts, we should not be surprised to find concepts such as *person*, *rational*, *meaning*, *agent causation*, *freewill*, *possibility*, *responsibility*, and *knowledge*—the bread and butter of much metaphysics. To serve as stepping stones, theories built using these concepts must look 'backwards' to science and 'forwards' to practical decisions. They must also be sensitive to the a priori deliverances of the non-derivational part of ethics. None of this requires that forensic metaphysics is in any sense a part of science; but nor is it evidentially dissociated from science.[23]

How, though, does this stepping stone model help us to deal with the threat posed by forensic metaphysical theories that appear to licence moral refutations of science? Suppose we are confronted with a metaphysical theory with both empirical roots and implausible moral implications. As an example, take the extended-mind thesis's apparently entailing that wiping someone's personal data is a kind of partial murder, an entailment that (let us suppose) we find hard to accept. What alternative do we have, on the stepping stone model, to treating such an implication (by hypothetical syllogism) as a moral strike against the empirical foundations on which the extended-mind thesis rests, contrary to working assumption 1?

We in fact have at least four alternatives.

(i) We could embrace the counterintuitive moral consequence, meaning the hypothetical syllogism would not be triggered.

This option is available because the consequence would be derived, and derived moral claims have no evidence-base beyond their derivability. Option (i) cannot be our only alternative, though, or the thought that brought down across-the-board naturalism—that sometimes a metaphysical theory's moral consequences *do* undermine its acceptability (see Section 7)—would also have purchase against this model of forensic metaphysics. Fortunately, the stepping stone model is compatible with further possibilities.

(ii) We could conclude that, after careful analysis, the metaphysical theory does not entail the untoward ethical conclusion after all.[24]
(iii) We could conclude that the metaphysical theory is not adequately rooted in good science after all.[25]
(iv) We could conclude that the metaphysical theory meshes poorly with non-derived ethical knowledge. [26]

Are four alternatives enough? Is any number enough? If the worry about Janus-faced forensic metaphysics had been that it *fails to rule out* moral refutations of science,

[23] The model has affinities with positions defended by others on different grounds, e.g. Sellars (1963), pp. 39–40; Jackson (1998), Chs. 5–6; Williams (2000); Paul (2012); Thomasson (2017).
[24] In the extended-mind case, for example, the entailment might fail because of an ambiguity: the conception of mind at work in the extended-mind thesis permits unconscious elements of self, whereas the notion of self at work in ethical contexts is tied to responsible agency and hence to a wholly conscious self.
[25] See e.g. Neil Levy's (2005) response to experiments by Benjamin Libet and others that are sometimes interpreted as showing that we lack freewill, a potential threat to ordinary assumptions about moral responsibility.
[26] For example, Lewis's rejection of 'pure' utilitarianism in Section 3 above.

alternative possibilities are beside the point: their existence does not make moral refutations of science impossible. That, however, was not the worry. We have already accepted that moral refutations of science cannot happen (i.e. working assumption 1, a brief rationale for which was given in note 12). Our task is to show that the stepping stone model does not *entail* their possibility. The existence of alternative possibilities undermines any such entailment.

In sum, the pluralist view divides metaphysical theories into two camps. Purely science-facing, non-forensic metaphysical theories, including analyses of explanatory paradigms, interpretations of particular theories, reflections on their mathematical underpinnings, etc., can be assimilated into science (albeit at a very abstract level). Given this, they are immune to moral refutation. Forensic metaphysical theories, on the other hand, are both science-facing and ethics-facing. They act as breakpoints in the long conditional paths linking science to its practical ethical consequences. This means they are, potentially anyway, susceptible to moral refutation—but not in a way that licences moral refutations of the science they presuppose.[27]

9 Is Metaphysics Immune to Moral Refutation?

The answer to this paper's title question turns out to be neither 'yes' nor 'no' but rather 'it all depends'. This is less disappointing than it sounds because we can now say *what* it all depends on, and why. As well as dispelling mystery and explaining divergences of opinion (e.g. the one between Heller and Tännsjö), this new knowledge can help us towards a verdict in particular cases.

What it depends on, in the first instance, is whether the metaphysical theory is best categorized as forensic or non-forensic. This will not always be easy to determine. In Section 7, we saw a couple of examples of philosophical theories the apparently forensic nature of which was unanticipated. There is an open question here as to how far we can *decide* how to categorize any given theory, and how far the appropriate category is a matter of discovery. While it may seem reasonable that someone should be free to stipulate their own ambitions, in the fashion of Papineau and Locke, we might also be tempted by the thought that such categorizations should sometimes be rejected because they push the theory in an unproductive direction. Either way, we can say that *if* the primary purpose of the metaphysical theory is to contribute to empirical enquiry, then it has moral immunity.

For forensic metaphysical theories, the picture is more complex. What makes it forensic is its serving as a resting post in the long and otherwise intractable conditionals linking science to the practical domain. When a piece of forensic metaphysics seems to have implausible moral consequences, we could be in any one of the four possible scenarios distinguished in Section 8. In (iii) and (iv), we should change the metaphysical theory; in (i) and (ii), we can make accommodations that do not require such a change. Deciding whether a metaphysical theory needs to be altered therefore depends on establishing which scenario one is in. This will often be just as hard as deciding whether the metaphysical theory is forensic in the first place. In practice, discussions of

[27] I am here ignoring the possibility of non-forensic *and* non-empirical metaphysical theories. Their existence could be accommodated within a pluralistic approach along lines suggested in note 14.

metaphysical concepts that arguably have an implicit or explicit forensic aspect, concepts such as *person* or *agency* for example, take place on ever-shifting sands, as we make definitional and other accommodations in an effort to reach an optimally coherent position. The question of which scenario we are in, and therefore the question of whether to revise the metaphysical theory on moral grounds, is one part of the attempt to find equilibrium. It will be answered only by addressing the particulars of the case.[28] All we can say for sure is that something should always give before an apparently implausible moral consequence of a metaphysical theory is recruited as evidence against one of the scientific hypotheses grounding that theory.

10 Conclusion: Lessons for Naturalism

The topic of the moral immunity of metaphysics was introduced with a question ('To what extent and in what respects should philosophy assimilate itself to science?') and a promise: that we can make progress on this question by considering moral immunity. So, what have we learned about philosophical naturalism in the course of thinking about the moral immunity of science and metaphysics?

The key lesson to draw is that we need not, indeed should not, insist on a one-size-fits-all approach to the status of metaphysics vis-à-vis science. It is perfectly possible for us to be insistent empiricists in some domains and conceptual jugglers, driven by a priori and practical considerations, in others.

Consideration of moral immunity also helps us to see that this pluralist approach is a more authentic species of naturalism than one that insists we always disregard non-empirical considerations. Naturalism in philosophy, characterized very broadly, is a matter of having a high regard for empirical enquiry as a defeasible route to knowledge, together with an open-minded willingness to go wherever this high regard takes us in our philosophy. Controversy takes root as soon as we try to pin this down further, but a healthy 'high regard' must exclude not only *distain* for science but also *sycophancy* towards it. Naturalism in philosophy implies, then, at least three dispositions:

> An active willingness to learn from science
> A concern not to pass oneself off as doing science when one is not
> A desire not to interfere inappropriately in science's ways

A naturalism that insists on all legitimate metaphysics being somehow assimilable to science would violate the second and third of these dispositions. Or rather, it would do so if we recognize that metaphysics can have legitimate forensic ambitions (as it seems we should). Many non-philosophical concepts, such as *noisy* and *dangerous*, are replete with evaluative significance, suiting them to what I called a stepping stone role, but undermining their use in science. The same seems to be true of many metaphysical concepts, such as *person*, *semantic/pragmatic*, *freewill*, *agent causation*, *meaning*, *rational*, and *possibility*.

[28] The kind of dancing involved is nicely illustrated by discussions of whether collectives are agents in a sense that is required for us to be able to hold them to account (e.g. Isaacs 2011).

Of course, supporters of a strong assimilationist reading of naturalism can respond by saying that such metaphysical concepts need to be re-engineered, divested of their empirically unhelpful 'legacy' features, including any forensic features. That is what 'naturalizing' philosophy requires, which is why naturalism is such a radical position. Biologists have refitted the ordinary concept *altruism* into a scientifically useful term of art devoid of moral overtones, so why can philosophers not do the same? But apart from the damage such refits do to the capacity of the revised notion to play its original and valuable forensic role, empirical success using this kind of conceptual exaptation is likely to be coincidental rather than the rule.

It is easier to respect all three dispositions if we adopt the pluralist view on which some metaphysics has a forensic element while some does not. Both types should be accountable to science. Even forensic metaphysics has scientific feet. But the ethics-facing aspect of forensic metaphysics means we should avoid seeing it as any kind of a contribution to or part of science. In practice, this means we need to find a way of ring-fencing science from moral refutation. This is possible if we accept that the purpose of some metaphysical theories is to mediate the long conditional inferences from science to practical decision-making. None of this bars us from describing *non*-forensic metaphysics as an extension of science, so long as other consideration in the naturalism debate call on us to do so.

Acknowledgements I am grateful to anonymous referees, Geraldine Coggins, Sean Cordell, Derek Matravers, Mark Pinder, Eduardo Garcia Ramirez, Joshua Thomas, and audiences at the Joint Session of the Aristotelian Society and Mind Association, the University of Reading, and Yeditepe University.

Compliance with Ethical Standards

Conflict of Interest The author declares that there is no conflict of interest.

References

Balashov, Y., & Janssen, M. (2003). Presentism and relativity. *The British Journal for the Philosophy of Science, 54*(2), 327–346.

Barber, A. (2013). Science's immunity to moral refutation. *Australasian Journal of Philosophy, 91*(4), 633–653.

Basl, J., & Coons, C. (2017). Ought to is: the puzzle of moral science. In R. Shafer-Landau (Ed.), *Oxford studies in metaethics 12* (pp. 160–186). Oxford: Oxford University Press.

Bentham, J. (1789). *An introduction to the principles of morals and legislation*. London: T. Payne and Son.

Carpenter, A. D. (2014). Ethics of substance. *Aristotelian Society Supplementary Volume, 88*(1), 145–167.

Chalmers, D. J., Manley, D., & Wasserman, R. (Eds.). (2009). *Metametaphysics*. Oxford: Oxford University Press.

Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis, 58*(1), 7–19.

Cohen, G. A. (2003). Facts and principles. *Philosophy & Public Affairs, 31*(3), 211–245.

Conee, E. (1999). Metaphysics and the morality of abortion. *Mind, 108*(432), 619–646.

Davidson, D. (1987). Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association, 60*(3), 441-458.
Fine, K. (2012). What is metaphysics? In T. E. Tahko (Ed.), *Contemporary Aristotelian metaphysics* (pp. 8–25). Cambridge University Press.
Forrester, J. W. (1984). Gentle murder, or the adverbial Samaritan. *Journal of Philosophy, 81*, 193–196.
Harman, G. (1977). *The nature of morality*. New York: Oxford University Press.
Hawley, K. (2006). Science as a guide to metaphysics? *Synthese, 149*(3), 451–470.
Heller, M. (2003). The immorality of modal realism, or: how I learned to stop worrying and let the children drown. *Philosophical Studies, 114*, 1–22.
Isaacs, T. (2011). *Moral responsibility in collective contexts*. New York: Oxford University Press.
Jackson, F. (1998). *From metaphysics to ethics: a defence of conceptual analysis*. Oxford: Oxford University Press.
Kant, I. (2002 [1785]) *Groundwork for the metaphysics of morals*. Allen Wood, ed. and trans. Newhaven: Yale University Press.
Kitcher, P. (1985). *Vaulting ambition: sociobiology and the quest for human nature*. Cambridge: MIT Press.
Ladyman, J., & Ross, D. (2007). *Every thing must go: metaphysics naturalized*. Oxford: Oxford University Press.
Levy, N. (2005). Libet's impossible demand. *Journal of Consciousness Studies, 12*(12), 67–76.
Levy, N. (2007). *Neuroethics*. Cambridge: Cambridge University Press.
Lewis, D. K. (1986a). *On the plurality of worlds*. Oxford: Blackwell.
Lewis, D. (1986b) Persons, morality, and tenselessness. *Philosophy and Phenomenological Research*, *47*(2), 305-309.
Locke, J. (1997 [1706]). In R. Woolhouse (Ed.), *An essay concerning human understanding* (5th ed.). London: Penguin.
Lowe, E. J. (2006). *The four-category ontology: a metaphysical foundation for natural science*. Oxford: Clarendon Press.
Millikan, R. G. (1996). On swampkinds. *Mind & Language, 11*(1), 103–117.
Newby, R. G., & Newby, D. E. (1995). The bell curve: another chapter in the continuing political economy of racism. *The American Behavioral Scientist, 39*(1), 12–24.
Oaklander, L. N. (1988). Delmas Lewis on persons and responsibility: a critique. *Philosophy Research Archives, 13*, 181–187.
Papineau, D. (1993). *Philosophical naturalism*. Cambridge: Blackwell.
Papineau, D. (2001). The status of teleosemantics, or how to stop worrying about Swampman. *Australasian Journal of Philosophy, 79*(2), 279–289.
Papineau, D. (2016). Naturalism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (winter 2016 edition)*.
Paul, L. A. (2012). Metaphysics as modeling: the handmaiden's tale. *Philosophical Studies, 160*, 1–29.
Rini, R. A. (2013). Feedback from moral philosophy to cognitive science. *Philosophical Psychology, 28*(4), 569–588.
Ross, D., Ladyman, J., & Kincaid, H. (Eds.). (2013). *Scientific metaphysics*. Oxford: Oxford University Press.
Saul, J. M. (2012). *Lying, misleading, and what is said: an exploration in philosophy of language and in ethics*. Oxford: Oxford University Press.
Sellars, W. (1963). Philosophy and the scientific image of man. In *Science, percetion and reality* (pp. 1–40). London: Routledge and Kegan Paul.
Stainton, R. M. (2016). Full-on stating. *Mind & Language, 31*(4), 395–413.
Stern, R. A. (1992). The relations between moral theory and metaphysics. *Proceedings of the Aristotelian Society, 92*, 143–159.
Stojanovic, I., ed. (2008). *The semantics/pragmatics distinction*: Special Issue of *Synthese. 165*(3).
Sturgeon, N. L. (1988). Moral explanations. In G. Sayre-McCord (Ed.), *Essays on moral realism* (pp. 229–255). Ithaca: Cornell University Press.
Szabo, Z. G. (2005). *Semantics versus pragmatics*. Oxford: Clarendon Press.
Tännsjö, T. (1987). The moral import of modal realism. *Theoria (Sweden), 53*, 87–96.
Thomasson, A. L. (2017). What can we do, when we do metaphysics? In G. d'Oro & S. Overgaard (Eds.), *Cambridge companion to philosophical methodology*. Cambridge: Cambridge University Press.
van Fraassen, B. C. (2002). *The empirical stance*. New Haven: Yale University Press.
Wallace, D. (2012). *The emergent multiverse: quantum theory according to the Everett interpretation*. Oxford: Oxford University Press.
Williams, B. (2000). Philosophy as a humanistic discipline. *Philosophy, 75*(294), 477–496.

Wilson, C. (1995). *The invisible world: early modern philosophy and the invention of the microscope*. Princeton: Princeton University Press.