15 Unbunking Arguments
A Case Study in Metaphysics and Cognitive Science

Christopher Frugé

A metainduction on the history of metaphysical debates suggests a healthy skepticism as to their eventual resolution. To some extent, of course, metainductions on most philosophical subjects lead to skepticism, but disputes in metaphysics seem particularly intractable. *When do two objects ever compose another?* The universalist says, "Always." The nihilist says, "Never." *Are the statue and the lump of clay constituting it two things or one?* The multithinger says, "Two!" The one-thinger says, "One!" *Is there an objective passage of time?* The A-theorist says, "Of course there is becoming. Time flows from past to present." The B-theorist says, "Of course there is no becoming. All times coexist in a timeless manner." At most one side can be right, but it is hard to see on what grounds one might decide between them. In this chapter, I outline three ways metaphysical disputes are prone to intractability and then suggest a way forward.

The first way in which resolutions are hard to come by is that similar problems arise for both sides of a metaphysical dispute. The second is that opposing metaphysical views are often equally descriptively powerful. The third is that metaphysical intuitions and beliefs have no source of independent confirmation. The realist metaphysician who wants to treat such disputes as substantive must give a reason to think progress can be made.

The most promising way to do so, I argue, is by giving an *unbunking argument*—the debunking argument's good twin. An unbunking argument for metaphysics would give reason to think that either all or a subset of metaphysical beliefs are caused by one or more processes that are reliable with respect to the relevant metaphysical truths, and it would do so without needing to assume any truths about the domain in question. The most promising sort of metaphysical unbunking argument, I suggest, relies on the *cross-domain strategy*. It argues that a process that is reliable in a given nonmetaphysical domain is also responsible for beliefs about a metaphysical domain and that we have good reason to think this process remains reliable

from one domain to the other. By way of demonstration, I examine an unsuccessful attempt to defend our modal beliefs due to Timothy Williamson (2008). I then formulate a more successful unbunking argument that relies on cognitive science.

15.1 INTRACTABILITY OF METAPHYSICAL DISPUTES

The first manner in which metaphysical questions are resistant to answers is that similar problems often arise for both sides of a metaphysical dispute.[1] Take the debate over whether there exist composite objects, those objects that have proper parts. On one side is the realist, who holds that composite objects exist. On the other side is the nihilist, who denies this. One of the nihilist's main arguments is that from causal overdetermination.[2] Let S be a collection of simple noncomposite particles arranged baseball-wise. Suppose S breaks a window. Now suppose that above and beyond the simples arranged baseball-wise there is in fact another object, the baseball, composed out of them. If so, the baseball is a sufficient cause of the window breaking, and so there are two sufficient causes of the window breaking: the baseball and the collection of simples arranged baseball-wise S. If there are indeed composite objects, this sort of causal overdetermination is rampant. Any effect of an event involving a collection of simples would be just as well explained by an event involving a composite object. On the supposition that there isn't rampant overdetermination, then there are no composites—at least assuming that composites can't be epiphenomenal and lack causal powers.

But as Karen Bennett notes, if overdetermination is a problem for the realist, then it is equally a problem for the nihilist (2009, 68–69). Take the collection of simples arranged baseball-wise S. Again, suppose S breaks a window. Now take another set of simples arranged baseball-wise S′, which is just S minus a single simple. S′ exists, hits the window, and is sufficient to break it, and so the window breaking is overdetermined. As before, this overdetermination is rampant. For any effect of an event involving some arrangement of simples, there is another event involving a different arrangement of simples that would also be sufficient to explain the effect. This is just one example of one problem in one dispute, but it is not unique. For many problems that plague one side of a metaphysical debate, there is an analogous problem that plagues the other.

The second issue with metaphysical disputes is that the respective views are often equally descriptively powerful in the sense that they can make all and only the same claims about the world, except with respect to the narrow range of facts about which they disagree. For instance, the compositional realist and the compositional nihilist

[1] The material of this paragraph and the next owes much to Karen Bennett (2009). See her paper for a more detailed discussion and further examples.

[2] Merricks 2001, 65–66. A wrinkle is that Merricks believes that persons exist, but we can ignore that fact for expository purposes.

can agree about everything, aside from putative composition facts. Take any world w. The realist and nihilist can describe this world in the exact same way, except that realist will claim that there are composition facts, whereas the nihilist won't. The realist can say there is a baseball that breaks the window. The nihilist can invoke plural quantification and say there are some simples that break the window. When two metaphysical views are equally descriptively powerful in the preceding sense, then the only direct counterexamples to either view would be one of the disputed putative facts. The only straightforward counterexamples to nihilism would be composite objects, but their existence is precisely what is under discussion. Of course, this sort of phenomenon is common to competing theories, but it is exacerbated in the metaphysical case given that the competing views are often empirically equivalent, and so there is no agreed-upon independent source of evidence that can provide a clear counterexample to either theory.

The third problem is that there is no independent way to corroborate our metaphysical beliefs and intuitions.[3] There is little external support for the verdicts of many metaphysical intuitions and beliefs, because there is no uncontroversial set of metaphysical facts to which we have access. This problem is exacerbated given that we have little initial reason to think we're particularly good intuitors of the joints of nature. As James Ladyman and Don Ross (2007) note:

> Proficiency in inferring the large-scale and small-scale structure of our immediate environment, or any features of parts of the universe distant from our ancestral stomping grounds, was of no relevance to our ancestors' reproductive fitness. Hence, there is no reason to imagine that our habitual intuitions and inferential responses are well designed for science or for metaphysics. (2)

In effect, Ladyman and Ross offer an evolutionary debunking argument of metaphysical beliefs and intuitions: evolution selected mechanisms that are responsible for our metaphysical beliefs and intuitions, but metaphysical truth does not enhance fitness and so there is no reason to think such mechanisms are reliable.

Perhaps the above problems leave the realist metaphysician with only the methodology of appealing to broad considerations of theoretical virtue and consilience with natural science—with little hope of ever coming to a single, indisputable metaphysical theory about a given domain.[4] However, in this chapter I explore whether there might be an argument that local metaphysical beliefs and intuitions are justified. In the next section, I examine the structure of negative debunking arguments before turning in section 15.3 to the structure of positive unbunking arguments. This latter sort of argument seems the best hope for realist metaphysicians who want to preserve some kind of justification for local metaphysical beliefs and intuitions.

[3] For the importance of external corroboration, see Weinberg 2007 and Sinnott-Armstrong 2006.

[4] For an argument that a multiplicity of metaphysical theories is a virtue, see Paul 2012.

15.2 DEBUNKING ARGUMENTS

Debunking arguments aim to show that we have no reason to think that our beliefs about a given domain are true. They rely on three premises. The first premise is *Objectivism*: we have beliefs about a domain that are true or false. This says that there are mind-independent facts about the domain in question, and it leaves open whether it is a fact that realism is correct or a fact that nihilism is correct. The second is the causal premise *Influence*: there is some set of causal processes that is primarily responsible for our beliefs about the domain, where the set may have one or more members. The third premise, *Off-track*, is epistemic: the set of processes do not reliably result in true beliefs. Thus, we have no reason to think our beliefs are true.[5]

We can schematize debunking arguments as follows:

1. Objectivism: beliefs about domain D are either true or false.
2. Influence: S's beliefs about D are due, dominantly, to set of causal processes X.
3. Off-track: set of causal processes X is unreliable with respect to D.
4. Skepticism: Therefore, S's beliefs about D are not justified.[6]

The following evolutionary debunking argument is an instance of this argument schema:

1. Objectivism: moral beliefs are either true or false.
2. Influence: S's moral beliefs are due, dominantly, to evolutionarily selected mechanisms.
3. On-track: evolutionarily selected mechanisms are unreliable when it comes to moral beliefs.
4. Nonskepticism: therefore, our moral beliefs are not justified.

I'll now offer some clarifications.

First clarification. Debunking arguments cannot merely show that one causal contributor to the beliefs about a domain is not reliable, for that is consistent with the *dominant* source or sources being reliable, and hence our beliefs being formed, overall, by a set of reliable processes. Instead, the causal premise must be that the dominant source or sources are not reliable, such that overall our beliefs about a domain are formed by a set of unreliable processes.

Second clarification. Most debunking arguments relativize reliability to a domain.[7] The evolutionary debunker of morality holds that evolution does not select for true

[5] See Goldman 1979 for the original defense of the view that justification and reliability are intimately related.

[6] For detailed discussions of the structure of debunking arguments, see Kahane 2011; Mason 2010; and Vavova 2014.

[7] Thanks to Alvin Goldman for drawing this feature to my attention.

moral beliefs, but presumably they would not want to say that evolution does not select for true beliefs about the trajectories of medium-sized physical objects. Thus, the debunker faces a generality problem of sorts. I will not solve any sort of generality problem here. However, it is worthwhile to note that the problem is less severe for debunkers, given that they are not offering a general theory of justification, but just an argument that we have no reason to think beliefs about a given domain are true. Insofar as we are just interested in beliefs about a given domain, we are licensed to restrict the measure of reliability to just that domain in question.

Third clarification. In this chapter, I connect reliability to justification, but even deniers of the link should find it important that beliefs about a domain are not reliably formed. Even someone who denies that reliability is a necessary condition on justification can agree that reliability is an epistemic virtue.

Fourth clarification. Crucially, debunking arguments are not metaphysical. They are epistemological. They do not directly argue for nihilism, that realist beliefs about the relevant domain are false. Instead, they argue for skepticism, that realist beliefs about the relevant domain aren't justified—whether or not they are true. In establishing Off-track, the debunker cannot assume antirealism without begging the question. This is disanalogous to showing, for instance, that our visual system can be mistaken. Our vision is prone to the Müller-Lyer illusion, where we experience two lines as being of different lengths when they are in fact of the same length. In this case, we can establish that the visual system is mistaken by appealing to independently verified facts, as when we measure the two lines with a ruler. Debunking arguments are particularly important when—as occurs in metaphysics—we have no uncontested domain of facts to which we can appeal in order to tell whether a given set of beliefs are justified or not.

15.3 UNBUNKING ARGUMENTS

In this section, I discuss the general structure of unbunking arguments before turning to a particular variation on that structure that relies on what I call the *cross-domain strategy*. This strategy argues from the reliability of a process or processes in one domain to the reliability of a process with respect to another domain. This strategy is particularly important when it comes to unbunking metaphysics, given that there is no uncontested set of metaphysical facts with which one can directly determine the reliability of a process.

15.3.1 The Structure of Unbunking

The basic structure of a debunking argument can be mirrored by a positive argument that leads to a nonskeptical conclusion. The first two premises remain the same. One assumes Objectivism about a domain, and one assumes that a particular set of causal processes is dominantly responsible for beliefs about that domain. The

only premise that changes is Off-track, and it becomes the epistemically positive *On-track*: the set of causal processes leading to belief is reliable with respect to the domain in question. The four clarifications mentioned previously still stand, pending changes making them relevant to this epistemically positive argument.

The following is the schema for unbunking arguments:

1. Objectivism: beliefs about domain D are either true or false.
2. Influence: S's beliefs about D are due, dominantly, to set of causal processes X.
3. On-track: set of causal processes X is reliable with respect to D.
4. Nonskepticism: therefore, our beliefs about D are justified.

To make things concrete, here's a straightforward example of an unbunking argument—which I don't necessarily endorse—that argues from the evolutionary origins of folk physics to its reliability. Folk physical beliefs are either true or false, and our folk physics is the result of evolutionarily selected mechanisms. Evolution selects for mechanisms that are useful to survival. It is useful to survival to accurately track the trajectories of medium-sized objects so that, for instance, we don't get hit by rocks. Therefore, we have good reason to think that our folk physics is largely accurate.

More formally:

1. Objectivism: folk physical beliefs are either true or false.
2. Influence: S's folk physical beliefs are due, dominantly, to evolutionarily selected mechanisms.
3. On-track: evolutionarily selected mechanisms are reliable when it comes to the physics of medium-sized objects.
4. Nonskepticism: therefore, our folk physical beliefs are justified.

Note that nowhere did the argument appeal to an independent set of facts with which one can check the deliverances of folk physics. The argument is not that one checked a thousand folk physical beliefs against cutting-edge theoretical physics and noticed that folk physics mostly got it right. Instead, the argument gave reason to think that folk physics is reliable independently of being able to directly determine its reliability.

If successful, a metaphysical unbunking argument would give us reason to think that our metaphysical beliefs and intuitions are justified. Unbunking arguments are uniquely suited to overcoming the problem that there is no independent source of evidence for metaphysical views outside of intuitions and beliefs. In just the same way that debunking arguments need not rely on truths about the domain in question, unbunking arguments need not either. Even if there is no way to independently confirm beliefs, we can still indirectly show they are generally justified.

15.3.2 Cross-Domain Strategy

How might one go about making a metaphysical unbunking argument? The first thing to do would be to argue for Objectivism. This I'll simply assume, and instead focus on how one might establish Influence and On-track. An immediate dead-end appears to be trying to come up with a wide-domain unbunking argument that shows that all metaphysical intuitions, regardless of stripe, are justified. Given the long history in metaphysics of unresolved disputes and the three problems listed in the first section, there isn't much reason to hope this could be successful. The more promising approach is to make a narrow domain unbunking argument, which argues for the more restricted claim that a certain class of metaphysical beliefs are justified.[8]

But there doesn't seem to be a straightforward unbunking argument available. For instance, appeal to evolutionary or cultural-historical causes of metaphysical beliefs is unable to help because there is no reason to think that either would select for beliefs that carve nature at the joints. True metaphysical beliefs are neither useful for survival, nor do they seem induced by cultural pressures. Therefore, the metaphysical unbunker needs to use a more indirect route.

I suggest unbunkers employ what I call the *cross-domain strategy*. They should argue (1) that a particular set of causal processes X′ is reliable with respect to some nonmetaphysical domain D′, where such reliability is uncontroversial; (2) that the same or a relevantly similar set of processes X is dominantly responsible for our beliefs about a metaphysical domain D; and (3) that D and D′ are relevantly similar so that we should expect X to be reliable with respect to D. Our justification for thinking our beliefs about D are reliable is inherited from our justification in thinking that our beliefs about D′ are reliable.

Here's a schematization of the cross-domain strategy style of unbunking argument, with the simplifying assumption that X and X′ are the same set of processes:

1. Objectivism: beliefs about domain D are either true or false.
2. Influence: S's beliefs about D are due, dominantly, to set of causal processes X.
   Lemma—cross-domain strategy:
   a. Set of processes X produces reliable beliefs about domain D′.
   b. Domain D′ and D are relevantly similar so if X is reliable with respect to D′, then X is reliable with respect to D.
3. On-track: X is reliable with respect to D.
4. Nonskepticism: therefore, our beliefs about D are justified.

Of course, the plausibility of an unbunking argument that uses the cross-domain strategy hinges on whether the two domains and processes are "relevantly" similar.

[8] This distinction maps onto Kahane's (2011) distinction between local and global debunking arguments.

I have nothing both general and useful to say about what makes for such similarity. Let's first just try to get some successful cases of the cross-domain strategy onboard before making an abstract analysis!

15.4 WILLIAMSON'S METAMETAPHYSICS

The contemporary philosopher who has articulated something closest to a metaphysical unbunking argument is Timothy Williamson. In his view, the methodology and forms of thinking used in philosophy are continuous with those used in other areas, including the sciences and ordinary life. There is no uniquely "philosophical" way of thinking. Instead, philosophy is just a specialization of ubiquitous ways of thinking, and so inherits its reliability from them. As Williamson (2008) himself puts it:

> One main theme of this book is that philosophical exceptionalism is false. . . . Although there are real methodological differences between philosophy and the other sciences, as actually practiced, they are less deep than is often supposed. In particular, so called intuitions are simply judgments (or dispositions to judgment); neither their content nor the cognitive basis on which they are made need be distinctively philosophical. In general, the methodology of much past and present philosophy consists in just the unusually systematic and unrelenting application of ways of thinking required over a vast range of non-philosophical inquiry. The philosophical applications inherit a moderate degree of reliability from the more general cognitive patterns they instantiate. (3).

While such a view, in broad outlines, seems correct, Williamson's implementation in the case of metaphysics falls short. He focuses on the metaphysics of modality and argues for its reliability in two stages. First, we are generally reliable in counterfactual thinking. Second, modal thinking about metaphysical possibility and necessity is just a special case of counterfactual thinking. Given these two claims, he argues, our modal thinking is generally reliable. In this section, I summarize the argument and then show how it fails.

As for the first step, Williamson primarily motivates the claim that ordinary counterfactual reasoning is generally reliable by invoking folk physics as an example. Take a rock rolling down a hill by a lake and hitting a bush and stopping. One can correctly judge that "If the bush had not been there, the rock would have ended in the lake" (142). We need not go and actually conduct an experiment, removing the bush and rolling the rock down the hill again. Instead, our folk physics allows us to make counterfactual judgments about different nonactual but possible trajectories of the rock.[9]

[9] Actually, folk physics is hopelessly inadequate when it comes to uncovering fundamental physics—superpositions, for one, are not countenanced at all in folk physics—and is notoriously prone to certain errors even in plotting the trajectories of macroscopic objects. For this last point, see McBeath et al. 2010 and McCloskey et al. 1983.

As for the next step, Williamson shows convincingly that a necessity claim is logically equivalent to a certain sort of counterfactual claim and that a possibility claim is logically equivalent to another sort of counterfactual claim (158). The following are the logical equivalences:

(Necessity) $\Box A$ if and only if $(\neg A > \bot)$

(Possibility) $\Diamond A$ if and only if $\neg(A > \bot)$

Necessity says that A is necessary if and only if its denial leads to a contradiction. A is necessary if and only if it is incoherent for A not to hold. Possibility says that A is possible if and only if it is not the case that A leads to a contradiction. A is possible if and only if it is not the case that it is incoherent for A to hold.

On the basis of these two claims, Williamson concludes that modal reasoning is generally reliable. But even granting that folk physics is generally reliable and that necessity and possibility claims are logically equivalent to certain counterfactual claims, his argument fails. To see why, it will be helpful to construe it as an unbunking argument:

1. Objectivism: metaphysical modal beliefs are either true or false.
2. Influence: S's modal beliefs are due, dominantly, to our ordinary capacity at handling counterfactuals.

Lemma—cross-domain strategy:

   a. Our ordinary capacity at handling counterfactuals is reliable with respect to everyday claims.
   b. Assessing everyday claims is not relevantly different from assessing claims of modal metaphysics.

3. On-track: our ordinary capacity at handling counterfactuals is reliable with respect to modality.
4. Nonskepticism: our modal beliefs are justified.

The argument fails because both Influence and On-track fail in this case.

As for Influence, all Williamson has shown is that necessity and possibility claims are logically equivalent to certain kinds of counterfactual claims. This does not show the same cognitive capacity is recruited to assess both modal claims and ordinary counterfactuals, and it does not even show that the cognitive capacity recruited to assess modal claims is relevantly similar to the one used for ordinary counterfactuals. Williamson says, "Whoever has what it takes to understand the counterfactual conditional and the elementary logical auxiliaries $\neg$ and $\bot$ has what it takes to understand possibility and necessity operators" (158). This might be so if one is appealing to explicit reasoning in a proof-theoretic system where the

inferential roles of >, ¬, ⊥, □, and ◊ are interdefined. But Williamson himself holds that counterfactual reasoning is not inferential and is instead imaginative (145–48). While logically one can move from one form of the statement to the other, the question at hand is whether our psychological capacity at assessing ordinary counterfactuals is used to assess modal claims. There is no reason to think that because two concepts are logically equivalent they are used identically in imagination, for they might be presented under different modes of presentation. The coextensionality claims of "□A if and only if (¬A > ⊥)" and "◊A if and only if ¬(A > ⊥)" might be Frege cases, where an agent does not realize that the left side is logically equivalent to the right side.[10] Because they conceptualize the left side differently from the right side, their manner of assessing claims when presented under one mode could very well be different than when presented under the other. Hence, there is no reason to think one and the same cognitive capacity used to assess counterfactuals is used to assess modal claims.

Even if Influence goes through, On-track still fails. It is plausible that our competency with ordinary counterfactuals might be fairly reliable with respect to certain sorts of domains, like plotting the trajectories of ordinary-sized objects, but not in others, like those pertaining to metaphysical claims. Even if we grant Williamson that the same cognitive capacity is recruited to assess ordinary counterfactuals as is recruited to assess metaphysical counterfactuals, there is no reason to think the latter are similar enough to the former for reliability to be inherited. Take Williamson's primary example of folk physics. At best, he has shown that we should be fairly reliable at assessing necessity and possibility claims as they pertain to trajectories of medium-sized physical objects. But there is no reason to think such reliability holds in judgments about, for instance, whether it is possible for physical duplicates of functioning humans to lack phenomenal consciousness. Indeed, the most current cognitive scientific view of folk physics is that it is a simulation engine, much like a graphics engine used to simulate physics in movies and video games (Battaglia et al. 2013). A graphics simulation engine has no way to represent counterfactuals involving colocation—let alone a host of other particularly metaphysical claims.

15.5 UNBUNKING METAPHYSICS USING COGNITIVE SCIENCE

This section is divided into two parts. In the first, I show how cognitive science can help in general to unbunk by giving an example of how work on the simulation

[10] Frege cases arise between logical equivalents all the time. Take the equivalence of "A v ~A" and "(B→C)→(((C→D)→B)→C)." Both are tautologies of propositional logic, but most of us have no explicit belief that the second one is a tautology, let alone that the two statements are equivalent. The difference in the modes of presentation of alethic modal claims and counterfactual claims is even more extreme.

theory of mindreading helps to unbunk our capacity to attribute mental states to other humans. In the second, I use results from cognitive science to support a cross-domain strategy unbunking argument for metaphysical beliefs about mutual supervenience. Often, cognitive science is used negatively against metaphysics, as when it is used to debunk metaphysical claims (for critical discussion, see Schaffer 2016), lower our credence in a realist view (Goldman 1989, 2015), or just generally make us aware of errors and biases leading to mistaken metaphysical commitments (Paul 2010). While this is a worthy enterprise, I want to explore another approach: using cognitive science to positively support metaphysical methodology.

15.5.1 Cognitive Science and Unbunking

In order for cognitive science to help unbunk it should accomplish three tasks. First, it needs to be able to identify processes dominantly responsible for our beliefs about a domain. Second, it needs to be able to help us assess their reliability. Third, it should allow us to discover when beliefs about one domain are produced by identical or at least similar processes that form beliefs about a second domain. This third task is required for unbunking arguments that rely on the cross-domain strategy, where one uses the reliability of a process with respect to one domain as evidence for the reliability of a process with respect to another. Simulation theory provides an example of cognitive science accomplishing all three tasks. I do not want to take a stand on whether the view is true, but simply want to use it as a clear case of cognitive science's ability to play a role in unbunking arguments.

The first task for cognitive science is to identify a set of one or more processes dominantly responsible for beliefs in a domain. According to simulation theory, one attributes mental states to another by imagining oneself in their situation.[11] One pretends to have the same mental states as the target of attribution and then feeds those states into one's own cognitive processes and attributes the resulting state or states to the target (Goldman 2012). A prime case is that of predicting another person's decision via imaginatively putting oneself in that person's place (Goldman 2006, 19–30). One pretends to have the same initial mental states, like beliefs and desires, and then one runs those beliefs and desires through one's own decision-making process. One attributes the resulting decision to the other person. For instance, while playing chess, I imaginatively put myself in my opponent's shoes and decide what move I would make in that situation. I then predict that my opponent will make that move.

The second task for cognitive science is to help us assess the reliability of belief-forming processes. These assessments can come in the negative and positive variety. As for the negative variety, simulation theorists have identified a certain sort of error in mindreading called "quarantine failure" (Goldman 2006, 29–30,

[11] In what follows, I rely heavily on the presentations in Goldman 2006 and 2012.

41–42). Quarantine failure occurs when one does not adequately isolate one's own nonpretend mental states while simulating another's mind. If these states are not shared by the target of attribution, they could undermine the accuracy of the simulation. For instance, when trying to predict my opponent's move in chess, if I do not adequately quarantine my beliefs about my own future move, then I might inaccurately simulate my opponent's move as being made—illicitly—in response to my own plans. The phenomenon known as "egocentric bias" provides evidence that quarantine failure is systematic. One instance of this bias is the "curse of knowledge," whereby people have trouble ignoring information they possess that the target of attribution doesn't (Camerer et al. 1989). This bias often makes it difficult for both children and adults to correctly attribute false beliefs to others (Birch and Bloom 2003; Birch and Bloom 2007). Simulation theorists, unfortunately, have done less work of the positive variety. However, cognitive science is certainly capable of doing work that uncovers how often we are reliable simulators and the conditions that promote such accuracy—even if that proves to be a less exciting research program.

The third task for cognitive science is to help discover when beliefs about two distinct domains are produced by similar or even identical processes. Simulation theory is proposed to account for a wide range of mental attributions, such as third-person attributions of sensations, emotions, propositional attitudes, and decisions as well as first-person attributions of future states of the same sort. Simulation theory holds that these different attributions are made using the same general routine. One feeds initial pretend states into one's own cognitive mechanism for the process to be simulated and one attributes the resulting mental state to the target of attribution. Simulation of each of these domains relies on a similar process, and so simulation theory in principle can provide support for a cross-domain strategy. Given that simulation is used to make attributions of sensations and emotions, we should expect roughly the same reliability with respect to the former domain as for the latter. Given that simulation is used to make attributions of both nonpropositional mental states and propositional attitudes, if we are very reliable in making the first sort we should expect to be at least somewhat reliable in making the second. Of course, simulation theorists can just directly test our reliability with respect to these domains and so the cross-domain strategy is not particularly needed here, but it is an important example of how cognitive science can support the strategy in general.

15.5.2 Cognitive Science Unbunking Metaphysics

I turn now to offering an actual unbunking argument for a certain class of metaphysical beliefs. Crucially, this argument relies on results from cognitive science to support an implementation of the cross-domain strategy. The unbunking argument I formulate revolves around identity claims whose justification comes from

mutual supervenience claims, which can be understood as special sorts of correlation claims. Work in cognitive science suggests that we are reliable at assessing such correlations.

Correlations often serve as evidence for identity claims. If two things are constantly observed to obtain together, then one potential explanation for this fact is that the two things are in fact one and the same thing. If Superman and Clark Kent always have the exact same haircut, one possible explanation is that Superman *is* Clark Kent. This sort of argument has been used to justify physicalism about the mental. As Brian McLaughlin (2010) formulates his argument for type physicalism, if we are justified in believing

> *the correlation thesis*: for any type of state of phenomenal consciousness C there is a type of physical state P such that it is true and counterfactual supporting that a being is in C if and only if the being is in P (267),

then we are justified in believing

> *type physicalism for phenomenal consciousness*: for every type of state of phenomenal consciousness C, there is a type of physical state P such that C = P (265)

on the basis of inference to the best explanation.

The correlation thesis says that C iff P is *counterfactual supporting*. The best way to understand this so that it supports the identity claim of C = P is that the thesis says C properties and P properties mutually supervene. Set of properties A supervenes on set of properties B if and only if a change in A requires a change in B (see McLaughlin and Bennett 2014). Therefore, A mutually supervenes on B if and only if a change in A requires a change in B and a change in B requires a change in A. The properties of Clark Kent mutually supervene on the properties of Superman. A change in Clark Kent's haircut requires a change in Superman's, and vice versa. The shape properties of a table supervene on its microphysical properties, but do not mutually supervene. A change in shape requires a change in microphysical properties, but not vice versa.

Although the mutual supervenience of the properties of two entities does not entail identity, identity entails mutual supervenience and so mutual supervenience can provide *evidence* for identity. Indeed, metaphysicians often appeal to considerations of mutual supervenience to argue for or against a metaphysical identification. The statue and the clay constituting it are often thought to be distinct because their properties do not mutually supervene. The statue can be destroyed by squishing it, but the clay will remain. The group agent is distinct from the mere aggregate of its human members because there can be a mere aggregate without any relevant unity to the group. The set of properties of the group agent mutually supervenes on the properties of an appropriate interrelation of its members, not the mere aggregate. Because beliefs about mutual supervenience play a role in generating metaphysical

commitments, unbunking mutual supervenience beliefs would unbunk an important class of metaphysical beliefs.

Mutual supervenience is a special sort of correlation. It is perfect correlation without possible exception. Therefore, the same processes by which we detect exceptionless correlations are those we use to detect mutual supervenience. If humans are good at assessing whether two items are perfectly correlated across possible scenarios, then they are good at assessing whether two items mutually supervene. The cross-domain strategy can exploit this fact. If we form reliable beliefs about correlations, then we should form reliable beliefs about mutual supervenience. The former can be tested with respect to domains that are uncontroversial and not overtly metaphysical, and so we can gain evidence as to our reliability with respect to the latter.

Putting this all together, here is an unbunking argument for beliefs about mutual supervenience:

1. Objectivism: mutual supervenience beliefs are either true or false.
2. Influence: S's beliefs about mutual supervenience are due, dominantly, to the set of causal processes by which we form beliefs about correlations.
   Lemma—cross-domain strategy:
   a. The set of processes by which we form beliefs about correlations is reliable with respect to these beliefs.
   b. Mutual supervenience beliefs are not relevantly different from correlation beliefs, so if the set of processes is reliable with respect to the latter, it is reliable with respect to the former.
3. On-track: the set of processes by which we form beliefs about mutual supervenience is reliable.
4. Nonskepticism: our beliefs about mutual supervenience are justified.

If successful, we have a metaphysical unbunking argument. Given that mutual supervenience claims can be used to justify metaphysical identity claims, we have reason to think local intuitions and beliefs about mutual supervenience are justified, as are identity claims based off them. The plausibility of the argument depends largely on premise 2 and subpremises *a* and *b*.

Though I argued for premise 2 and subpremise *b* above, there has not been cognitive scientific work directly addressing them. However, they are claims clearly amenable to cognitive scientific research. The general methods that cognitive scientists use to get at judgments, for example, of causation and responsibility can be used to gather data about judgments of mutual supervenience. There is nothing that would keep researchers from designing experiments testing how people judge whether two items obtain and change together. Moreover, scientific theorizing and fMRI results can provide evidence for whether the same process used to form beliefs about correlations in general is responsible for beliefs about mutual supervenience in

particular—as, for instance, when fMRI results suggest that past, future, and counterfactual thinking involve many common brain processes (Van Hoeck et al. 2012).

Fortunately, there has been a lot of cognitive scientific research pertaining to subpremise *a*. Cognitive scientific study of our ability to make judgments of correlation often goes under the heading of assessments of "covariation" or "contingency." The role of beliefs about covariation has been studied in a wide range of areas in psychology, but what is relevant for my particular unbunking argument is our ability to detect and predict correlations. Studies have tended to provide evidence that we are good detectors in general, though prone to certain biases and errors.

As one example, Well et al. (1988) presented subjects with three sets of sixty pairs of numbers of correlation .9, .6, and .1 respectively. One group had both numbers presented simultaneously, and after all sixty pairs were presented they were asked to judge the "strength of the relationship." Another group was forced to make predictions of the last thirty pairs in each set of the second of the numbers on the basis of the first, and after all sixty pairs were presented they were also asked to judge the "strength of the relationship." Neither group of subjects judged the strength of the relationship in a manner corresponding to standard statistical measures of correlation, but the prediction group was quite good at predicting one variable from another in a manner sensitive to the degree of correlation. In a related experiment, Malmi (1986) encouraged subjects to rely on "intuitive" implicit processes to make covariation judgments and found that people were quite good at estimating degree of correlation. In a series of experiments, subjects were presented with pairs of numbers, pairs of lines of differing lengths, and pairs of words and lines. Malmi encouraged "intuitive" strategies of assessing correlations by presenting the pairs for one second each. He found that people were in general quite good at estimating the sign and degree of correlation.

A survey of research on covariation by Alloy and Tabachnik (1986, 119–23) concluded that humans are prone to errors and biases in detection of covariation but that under many conditions we are quite accurate. Accuracy is promoted if the correlated items

are positively correlated
are neutral with respect to rewards
are not associated with success
are not associated with features characteristic of situations requiring skill
are preceded by something that makes randomness a salient possibility

Often these conditions are met when we encounter mutually supervening items.

Two issues with drawing conclusions about our ability to track mutual supervenience from such studies should be addressed. The first is that the studies testing our ability to detect correlations have only tested subjects' ability to detect nonmodal frequencies. They have neither presented the subjects with modal

information nor asked them to make modal judgments. The second issue is that the studies give artificial tasks different from those that subserve our mutual supervenience beliefs. Judging the correlation between two numbers and the lengths of two lines is different from judging whether the team mutually supervenes on the players under some teammate relation.

As for the first issue, the studies give reason to think that if humans are good at picking up nonmodal frequency information, then they would be good at picking up modal frequency information. Moreover, cognitive scientists can devise studies to get at our modal correlation judgments—as they do when they devise studies getting at our modal judgments about causation. As for the second issue, the artificiality of the tasks gives us reason to think humans would be even better at detecting non-artificially constructed correlations. And, as before, testing less artificial judgments as to correlations is something amenable to cognitive scientific research.

It is worthwhile to remember that studies showing positive results have given subjects items that aren't perfectly correlated and have shown them relatively few examples, generally a hundred or fewer. The relevant sorts of correlations for mutual supervenience claims are those that are positive, without possible exception, and, often, ubiquitous. Many of the cases relied on in metaphysical theories are of common objects, such as clay statues. Most people see more than a hundred statues in their lives, and certainly many more materially constituted objects. There is good reason to think that if humans are good at detecting imperfect correlations with only a few examples, then we would be even better at detecting and predicting perfectly correlated items we encounter much more frequently.

Given the cognitive scientific research, it appears that beliefs about metaphysical mutual supervenience are unbunked.

15.6 CONCLUSION

In the first part of the chapter I raised three issues with metaphysical practice that suggested the need for an unbunking argument for metaphysical beliefs. The first is that similar problems often arise for competing sides in a dispute. The second is that metaphysical views can make all the same claims as other metaphysical views, except for the narrow range of propositions about which they disagree. The third is that there are good reasons to think human cognition isn't particularly well suited to uncover the fundamental joints of nature. I suggested that the realist—and optimistic—metaphysician could justify certain classes of metaphysical beliefs by appeal to unbunking arguments, particularly those that rely on the cross-domain strategy. Of course, merely having justified metaphysical beliefs about a domain doesn't guarantee that different metaphysical views about that domain won't face similar problems, nor certainly make it such that different views have differing levels of descriptive power. But it does give us reason to think that we are good enough

carvers of nature, at least that realm of nature for which we have an unbunking argument.

Unbunking arguments need not be limited to metaphysics. If a successful one can be found for metaphysics—the philosophical subject seemingly most prone to irresolvable disputes—then we should remain optimistic enough that unbunking arguments can be found for other philosophical subjects.

ACKNOWLEDGMENTS

For their helpful comments and discussion, I thank Carolina Flores, Adam Gibbons, Kyle Landrum, Tyler John, John McCoy, and Brian McLaughlin. I want to especially thank Alvin Goldman, who gave me many detailed comments, as well as Danny Forman, who pointed out my nascent views had structural similarities to a debunking argument.

REFERENCES

Alloy, L., and Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review 91* (1): 112–49.

Battaglia, P. W., Hamrick, J. B., and Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences of the United States of America 110* (45): 18327–32.

Bennett, K. (2009). Composition, colocation, and metaontology. In D. Chalmers, D. Manley, and R. Wasserman, eds., *Metametaphysics: New Essays on the Foundations of Ontology*. New York: Oxford University Press, 38–77.

Birch, S. A. J., and Bloom, P. (2003). Children are cursed: An asymmetric bias in mental-state attribution. *Psychological Science 14* (3): 283–86.

Birch, S. A. J., and Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science 18* (5): 382–86.

Camerer, C., Loewenstein, G., and Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy 97* (5): 1232–54.

Goldman, A. (1979). What is justified belief? In George Pappas, ed., *Justification and Knowledge*. Dordrecht: D. Reidel, 1–25.

Goldman, A. (1989). Metaphysics, mind, and mental science. *Philosophical Topics 17* (1): 131–45.

Goldman, A. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. New York: Oxford University Press.

Goldman, A. (2012). Theory of mind. In E. Margolis, R. Samuels, and S. Stich, eds., *The Oxford Handbook of Philosophy and Cognitive Science*. New York: Oxford University Press, 402–24.

Goldman, A. (2015). Naturalizing metaphysics with the help of cognitive science. In K. Bennett and D. Zimmerman, eds., *Oxford Studies in Metaphysics*, vol. 9. New York: Oxford University Press, 171–216.

Kahane, G. (2011). Evolutionary debunking arguments. *Noûs 45* (1): 103–25.

Ladyman, J., and Ross, D., with Spurrett, D., and Collier, J. (2007). *Every Thing Must Go: Metaphysics Naturalized*. New York: Oxford University Press.

Malmi, R. (1986). Intuitive covariation estimation. *Memory & Cognition 14* (6): 501–8.

Mason, K. (2010). Debunking arguments and the genealogy of religion and morality. *Philosophy Compass 5* (9): 770–78.

McBeath, M. K., Brimhall, S. E., Miller, T. S., and Holloway, S. R. (2010). The naïve physics curvilinear impetus bias does not occur for locomotion. *Journal of Vision 10* (7): 1021.

McCloskey, M., Washburn, A., and Felch, L. Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition 9* (4): 636–49.

McLaughlin, B. (2010). Consciousness, type physicalism, and inference to the best explanation. *Philosophical Issues 20* (1): 266–304.

McLaughlin, B., and Bennett, K. Supervenience. In E. N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*. Spring 2018 ed. https://plato.stanford.edu/archives/spr2018/entries/supervenience/.

Merricks, T. (2001). *Objects and Persons*. Oxford: Clarendon Press.

Paul, L. A. (2010). A new role for experimental work in metaphysics. *Review of Philosophy and Psychology 1* (3): 461–76.

Paul, L. A. (2012). Metaphysics as modeling: The handmaiden's tale. *Philosophical Studies 160* (1): 1–29.

Schaffer, J. (2016). Cognitive science and metaphysics: Partners in debunking. In B. McLaughlin and H. Kornblith, eds., *Goldman and His Critics*. Malden, MA: Wiley-Blackwell, 337–68.

Sinnott-Armstrong, W. (2006). Moral intuitionism meets empirical psychology. In T. Horgan and M. Timmons, eds., *Metaethics after Moore*. New York: Oxford University Press, 339–65.

Van Hoeck, N., Ma, N., Ampe, L., Baetens, K., Vandekerckhove, M., and Van Overwalle, F. (2012). Counterfactual thinking: An fMRI study on changing the past for a better future. *Social Cognitive and Affective Neuroscience 8* (5): 556–64.

Vavova, K. (2014). Evolutionary debunking of moral realism. *Philosophy Compass 10* (2): 1–13.

Weinberg, J. (2007). How to challenge intuitions empirically without risking skepticism. *Midwest Studies in Philosophy 31* (1): 318–43.

Well, A. D., Boyce, S. J., Morris, R. K., Shinjo, M., and Chumbley, J. I. (1988). Prediction and judgment as indicators of sensitivity to covariation of continuous variables. *Memory & Cognition 16* (3): 271–80.

Williamson, T. (2008). *The Philosophy of Philosophy*. Malden, MA: Wiley-Blackwell.