sure that I have succeeded in even that more modest task. I myself find physicalism to be sufficiently nonthreatening to responsibility that I have difficulty gauging how threatening others find it to be.[85]

[85] I have some limited data here. Just before I was to give a version of this chapter at the University of Pennsylvania, Center for Neuroscience and Society, my wife and usual philosophical collaborator, Heidi Hurd, e-mailed me her views on the separation of the determinist skeptic from the reductionist skeptic:

Just say, "there ARE two different debates here about TWO different things that persons standardly conflate" very loudly and repetitively (with that impatient, slightly disdainful tone in your voice that you always get when you are on philosophically thin ice), and people will be sure that they're the idiots when they can't see why the distinction you are drawing makes any difference!!

3

Non-Eliminative Reductionism: Not the Theory of Mind Some Responsibility Theorists Want, but the One They Need

KATRINA SIFFERD*

In this chapter I will argue that the criminal law is most compatible with a specific theory regarding the mind–body relationship – non-eliminative reductionism. Criminal responsibility rests upon mental causation: a defendant is found criminally responsible for an act where she possesses certain culpable mental states that are causally related to criminal harm in the appropriate way. If we assume the widely accepted positon of ontological physicalism, which holds that only one sort of thing exists in the world – physical stuff – non-eliminative reductionism about the mind offers the most plausible account of the full-fledged mental causation criminal responsibility requires. Other theories of the mind–body relationship, including eliminativism and non-reductive physicalism, threaten criminal responsibility because they do not offer satisfactory accounts of mental causation. Eliminativism, as the name implies, eliminates the mental or is skeptical that it can do the causal work necessary to responsibility; and non-reductive theories disconnect the mental from the physical/casual world such that the mental can no longer have reliable causal effects.

Mental Causation and the Criminal Law

To be criminally responsible, a defendant must be found to possess certain mental states at the time a crime was committed, and these mental states must be causally related to criminal harm for which the

* The author would like to thank Ty Fagan and Bill Hirstein for their very helpful comments on an earlier draft of this chapter, and Michael S. Moore and Stephen Morse for inspiring and critically engaging with the ideas presented here. Thank you also to David Papineau, who taught me to be a physicalist.

defendant has been arrested. This means that the criminal law places mental states in a privileged role in the explanation of human action, where such states are seen as the source or cause of behavior.[1] The theory of human psychology grounding the criminal law is a folk, or commonsense, psychological theory.[2] This means the criminal law trades in, and has codified, the psychological theory that most human beings naturally utilize to attribute psychological states to themselves and others when they attempt to explain or understand human behavior.[3] The folk theory of psychology, also called our "theory of mind," emerges very early in human development and uses behavioral cues, as well as knowledge about persons' character and preferences, etc., to attribute mental states to others[4]. So, even fairly young children pass the "false belief" test, which is a measure of whether a subject can attribute specific belief states that they themselves do not hold: a child as young as three can figure out that a doll who last viewed a toy in a chest will look for it in the chest, even though the test administrator moved the toy out of the chest when the doll was "away."[5]

[1] On this, see: Morse, S. "Neuroscience and the Future of Personhood and Responsibility" in J. W. Rosen ed. *Constitution 3.0: Freedom and Technological Change* (Washington, DC: Brookings, 2011); Morse, S. "Inevitable Mens Rea" (2003) 27 *Harvard Journal of Law & Public Policy*; Morse, S. "Criminal Responsibility and the Disappearing Person" (2007) 28 *Cardozo Law Review*, 2545–2575; and Sifferd, K. L. "In Defense of the Use of Commonsense Psychology in the Criminal Law" (2006) 25 *Law and Philosophy*, 571–612. Prominent legal scholar Stephen Morse claims that our ordinary understanding of human behavior posits that "virtually all actions for which agents deserve to be praised, blamed, rewarded, or punished are the product of mental causation" (Morse, ibid, 2011, at 530).

[2] See Sifferd, ibid.

[3] One group of persons who may not have this commonsense "theory of mind" necessary to attribute mental states to others is autistics. See Baron-Cohen, S. *Mindblindness: An Essay on Autism and Theory of Mind: Learning, Development and Conceptual Change* (Cambridge, MA: MIT Press, 1995). For more information on folk psychology, see: Fodor, J. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (Cambridge, MA: MIT Press, 1987); Jackson, F. and P. Pettit "In Defense of Folk Psychology" (1990) 59 *Philosophical Studies*, 31–54; Morse, S. "Determinism and the Death of Folk Psychology: Two Challenges to Responsibility from Neuroscience." (2008) 9(1) *Minnesota Journal of Law, Science and Technology*, 1–36.

[4] Gopnick, A. and H. M. Wellman, "Why the Child's Theory of Mind Really Is a Theory" (1992) 7 *Mind & Language*, 145–171.

[5] Wimmer, H. and J. Permer, "Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception" (1983) 13 *Cognition*, 103–128.

The U.S. Model Penal Code asks courts to sort the folk psychological mental states assigned to criminal defendants into four defined categories: (1) purposeful, (2) knowing, (3) reckless, and (4) negligent.[6] For example, a homicide might be committed where the death was caused "purposely" (e.g., where the shot was fired for the purpose of killing the shooter's correctly identified target); "knowingly" (e.g., where the shooter actually intended to scare the victim, but knew there was a chance she would be killed); or "recklessly" (e.g., where the defendant should have known of the high risk that someone would be killed when he shot into the crowd). At trial, judges and jurors hear evidence of the defendant's behavior, which the prosecution hopes will establish both that the defendant had the opportunity to commit the crime and that he possessed the requisite mental states regarding the criminal harm caused to be found guilty. The prosecution may present evidence that the defendant was seen intentionally pointing the gun at the victim, or made prior threats against the victim; and evidence of motive (e.g., reasons the defendant would benefit from the victim's death) may be offered. Different levels of punishment may be warranted depending upon the level of mental state assigned: a purposeful homicide will result in more serious punishment than a reckless one.

Because the criminal law is couched in folk psychological terms, scientific evidence of a defendant's mental state, such as a medical diagnosis, evidence of a mental disorder, or fMRI or PET scan evidence, should only be presented to the court if such evidence undermines or supports attribution of a commonsense mental state at issue or if it supports an affirmative defense (e.g., mental incapacity or legal insanity), also couched in folk terms.[7] For example, evidence of a brain tumor is relevant to a defendant's guilt only if it indicates he did not possess the intentional mental states required under the law or to help establish that some excuse, such as diminished mental capacity, existed at the time of the crime.

For a defendant to be found guilty the prosecution must also prove a causal link between his culpable mental state and the criminal harm for

[6] (1985). Model Penal Code and Commentaries (Official Draft and Revised Comments). The Model Penal Codes was drafted to be used a model for state criminal codes, to increase the consistency of codes across the various U.S. states.

[7] One exception to this general rule may be the requirement of a "mental disease or defect" by the M'Naghten rule for legal insanity, which may require a mental disease be presented by an expert witness in terms of a scientific psychological theory.

which he has been arrested.[8] Certain very odd cases highlight the need to establish, and not assume, this link. Imagine a case where a defendant, Mike, intended to kill his neighbor, and Mike's behavior actually caused the neighbor's death. However, Mike had no intention of killing the neighbor at the time when he caused his neighbor's death: Mike set his own backyard shed alight, having no clue his neighbor was in there; indeed, Mike intended to kill his neighbor later that evening by poisoning him. In this case the actual death of Mike's neighbor was not causally related to Mike's intention to kill him, and Mike cannot be held responsible for murder (although he may have been negligent in the way he set fire to his shed).

The link between a defendant's mental state and the criminal harm must thus be thought of as a separate component of guilt. However, this link can be established in nonlinear ways, as happens with cases of "transferred intent." In a typical homicide case, the prosecution aims to prove that the defendant desired the victim's death and held the belief that some course of action would cause this death; and that the defendant then performed this action, which did indeed cause the victim to die. However, imagine a case where I intend to kill Anya, but my hand shakes so much that I kill Xander (who was standing next to her) instead. Under the doctrine of transferred intent, the court may transfer my purposeful intent to kill Anya to the criminal harm of Xander's death. In this atypical case, the criminal harm is causally linked to my culpable mental states because my desire to kill caused me to pull the trigger, but the law is willing to transfer the intent from the intended target for another. The situation becomes more difficult if Anya sees me pointing the gun at her and then pushes Xander into the path of the bullet, because my culpable intent to kill may no longer be considered the proximate cause of Xander's death. In this case Anya may also be considered culpable for Xander's death: if the court finds Anya acted knowing it was likely Xander would die, it seems clear her mental states are causally linked to Xander's death.

These examples are offered to emphasize how crucial mental state causation is to attributing criminal responsibility to defendants and thus, also to criminal punishment. It should be obvious that if it turned out to

[8] Criminal guilt typically requires that the defendant's behavior is the proximate cause of the criminal harm: this means that without the defendant's behavior, the criminal harm would not have occurred.

be the case that the mental states the criminal law seeks to attribute to defendants weren't *real* – if either (1) commonsense mental state concepts don't reliably refer to states or events within defendants or (2) such states or events fail to serve as a causal factor in criminal harm, then attributions of criminal responsibility by the criminal law are unjustified. As prominent legal scholar Stephen Morse notes, the law will be "fundamentally challenged" if it is shown that the commonsense psychology the law depends upon "is wrong and we are not the type of creatures for whom mental states are causally effective."[9] To put it bluntly: if mental causation isn't true, we are putting people in prison for no principled reason. Anyone hoping to vindicate the responsibility attributions made via the criminal law must seek a theory of the mind-body relationship that preserves the reality of commonsense mental states and their causal powers.

In the past 50 years or so there has been a lively philosophical debate regarding whether such a theory can be found. Critics of commonsense psychology, such as Paul and Patricia Churchland, expressed doubts about any theory of the mind–body relationship because they claim the mind (couched in folk terms) doesn't exist. Commonsense psychology, they argued, is unlikely to be true because the theory described entities in our heads (mental states), and these entities could not be directly perceived via the human senses – they were formed at a time when we at best had indirect access to the properties of brain entities or processes.[10] On the other hand, "realists" about commonsense mental states such as Jerry Fodor argued that the truth of commonsense psychology is evident from its predictive power. Even though the predictive power of commonsense psychology isn't completely accurate, it does seem indispensable to human existence, precisely because of how well it works.[11] Although commonsense explanations seem to work ceteris paribus, meaning that they aren't without exception, these ceteris paribus phrases are necessary to form mental state generalizations in all of the special

[9] Morse, 2011, 534.

[10] See Churchland, P. "Eliminativist Materialism and the Propositional Attitudes" (1981) *Journal of Philosophy*, 78; Churchland, P. *Neurophilosophy: Toward a Unified Science of the Mind-Brain* (Cambridge MA: MIT Press, 1986); Churchland, P. and P. S. Churchland *On the Contrary: Critical Essays, 1987–1997* (Cambridge, MA: MIT Press, 1999).

[11] Fodor, J. "Fodor's Guide to Mental Representation: The Intelligent Auntie's Vade-Mecum" (1985) 373 *Mind*, 76–100.

sciences, and we don't doubt the existence of the items studied by economics or political science.[12]

Recently the eliminative position has been reworked by some scholars to claim not that commonsense psychological concepts fail to refer, but that folk psychology assumes a particular type of mental causation that is false. For example, philosophers Derk Pereboom and Gregg Caruso argue that legal responsibility assumes a freedom of will or type of agential control that human beings do not actually have.[13] They conclude that because we do not have the required level of control over our actions, they cannot be considered free, and thus no one is ever responsible for their acts; similarly, no one ever really deserves criminal punishment.[14] Caruso concludes that dangerous actors should be "quarantined" instead.[15] Many compatibilists about free will disagree, of course, and some legal scholars have argued specifically that the criminal law is compatible with the truth of determinism and its descendants.[16]

Even for those who believe that commonsense mental state concepts do truly refer, and that folk psychology describes a causal process that can ground responsibility, the task remains to posit a relationship between mental states couched in commonsense terms and their physical nature such that mental causation is possible. The nature of this relationship depends crucially on a theorist's ontological view of world.

[12] Fodor, J. "Special Sciences (or the Disunity of Science as a Working Hypothesis)" (1974) 28 *Synthese*, 97–115.

[13] See, Caruso, G. "Free Will Skepticism and Criminal Behavior: A Public Health-Quarantine Model" (2016) 32 *Southwest Philosophy Review* and Pereboom, D. "Free Will Skepticism and Criminal Punishment" in T. Nadelhoffer ed., *The Future of Punishment* (New York: Oxford University Press, 2013), 49–78.

[14] See Pereboom, D. and G. Caruso "Hard-Incompatibilist Existentialism: Neuroscience, Punishment, and Meaning in Life" in G. Caruso and O. Flanagan (eds.), *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience* (Oxford: Oxford University Press, 2018).

[15] Caruso, 2016.

[16] See Morse, S. "Lost in Translation?: An Essay on Law and Neuroscience" (2011) *Law and Neuroscience, Current Legal Issues*, 13 and Morse, S. "Neuroscience, Free Will, and Responsibility" in W. Glannon ed., *Free Will and the Brain: Neuroscientific, Philosophical, and Legal Perspectives* (Cambridge: Cambridge University Press, 2015). Many now think universal determinism is likely to be false because quantum mechanics indicates that much of what happens in the universe is indeterminate (Pereboom and Caruso, 2018 at 9). But of course the notion that our actions may result from indeterminate events makes agential control even less likely than the notion that our actions are related to determined events.

For scholars committed to physicalism – the idea that only one thing exists in the world, and it is physical stuff – the question posed is how to understand the way in which different epistemic entities (entities described under a scientific theory, and entities described under the folk psychological theory) are related.[17] If everything that exists is physical stuff, mental states are physical stuff. The question is what sort of stuff they are, and how we are to understand the relationship between the laws that govern mental states under a physical description, versus the rational relationships they seem to exhibit as folk psychological explanations.

To sum up this section: criminal verdicts crucially depend upon *mental causation*, the notion that our mental states can cause things to happen in the world. I will argue later that only certain theories regarding the relationship between mental states and physical states (such as brain states) preserve mental causation. Our best candidate theory for preserving mental causation, and the practice of attributing criminal responsibility to defendants, and punishing them, is a non-eliminative reductive theory of mental states. But first, I will argue in the next section that the two forms of eliminativism mentioned earlier are false.

I will only discuss theories that claim to be compatible with physicalism. This is because I think the problems ontological dualism poses for mental causation are well known and have been dealt with thoroughly elsewhere. I will not attempt to offer a complete account of any theory discussed; but instead hope to highlight the difficulties each theory poses for mental causation, with the idea that I might show my favored theory, non-eliminative reductionism, is the best theory to preserve mental causation, and thus the best theory to ground responsibility attributions made in the criminal law.

[17] As Ney notes: "Physicalism is, if not a received view, at least a starting point in much of philosophy today. There are those who dissent, but the view that physics, perhaps even a physics of the not-too-distant future, may provide an exhaustive guide to the fundamental constituents of our world, a view rejecting fundamental mental and biological phenomena, is widely adopted as a place to begin debates about fundamental ontology, the nature of mind, the special sciences, and even the status of ethics and normativity." Ney, A. "Microphysical Causation and the Case for Physicalism" (2016) 57 *Analytic Philosophy*, 141–164. For an argument in support of physicalism, see Papineau, D. *Philosophical Naturalism* (Oxford: Blackwell, 1993).

Elminativism

Eliminativism about Commonsense Psychological Concepts

If eliminativism about mental states is true, criminal verdicts generated by looking for particular mental states in defendants, causally related to criminal harm, would be no different than criminal verdicts generated via phrenology or by throwing the defendant into a vat of water to see if she sinks or swims. This brand of eliminativism could not possibly preserve mental causation because it does not preserve commonsense mental concepts. If mental state concepts must be eliminated, attribution of criminal responsibility and punishment ought to be suspended.

Paul and Patricia Churchland, the most well-known defenders of eliminativism in the 1980s and 1990s, offer several arguments that commonsense psychology fails to refer to real entities in the world and must be eliminated. The Churchlands have claimed commonsense psychology (1) has not "shown the expansion and developmental fertility one expects from a true theory"; (2) shows no signs of being smoothly integrable with "... the emerging synthesis of the several physical, chemical, biological, physiological, and neurocomputational sciences; and (3) fails to "... to explain a considerable variety of central psychological phenomena" like mental illness, memory, intelligence differences, and the different forms of learning. A viable theory, they argued, would show more progress, be more easily integrated with scientific theories, and would not have such explanatory gaps.[18]

However, close examination of these arguments reveals they do not provide support for the idea that commonsense psychology posits concepts that fail to refer to anything in the world. Indeed, there seems to be good evidence that commonsense psychological concepts do reliably refer to states or events in our heads: as indicated earlier, such concepts actually work very well to explain our behavior.[19] If mental state concepts did not refer to fairly stable entities in our heads, that in turn had fairly reliable causal effects, how could such concepts be used so successfully to predict human behavior? For example, if thirst did not exist and thus failed to have a stable causal identity, how might we accurately predict – all things being equal – that persons who are thirsty will go get a drink?

[18] See Churchland and Churchland, 1999.

[19] Fodor, J. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (Cambridge, MA: MIT Press, 1987).

In sum, as Fodor has noted, one big reason to think commonsense psychology isn't an utterly false theory is that it *works so well*.[20]

Another way to judge the veracity of commonsense psychological concepts is to look to see whether their referents are indeed emerging from progress made in scientific psychology. Does our best current scientific psychology appear to vindicate or debunk the notion that perceptual states like pain and thirst, emotional states like anger, and informational states like the belief that it is raining, have existence as brain states? In the very least, evidence generated by cognitive science and neuroscience indicate that mental states such as pain, anger, thirst, and beliefs do indeed exist and are instantiated in the brain.[21] The search for the functional identity of such commonsense mental states seems to be well underway; some might even agree that it is progressing nicely.

Given this, it isn't clear why the Churchlands seemed convinced that commonsense psychology shows no signs of being integrable with the hard sciences. Insofar as there is progress in neuroscience, this progress is made via discovering the referents of commonsense mental concepts. This is because the primary way to search for functional capabilities – or dysfunction – in the brain is by attempting to locate the referents of commonsense concepts in the brain. Neuroscientists use commonsense concepts to describe a cognitive capacity such as facial recognition (a capacity that can be understood as the ability to form beliefs that something is a face) before they go looking for that capacity in the brain. They then seek the capacity or incapacity by giving subjects cognitive tasks such as identifying faces – and indicating they see a face using commonsense terms – while, for example, the subjects are in an fMRI machine ("Yes, I see a face now," and "Nope, I don't see a face.").

Further, regarding the Churchlands' claim that an "emerging synthesis" regarding the nature of the brain is currently under construction by

[20] Ibid. at 3–4.

[21] To offer just a few examples: Egan, G. et al. "Neural Correlates of the Consciousness of the Emergence of Thirst" (2003) 100 *Proceedings of the National Academy of Sciences of the United States of America*, 15241–15246; Denson, E. F. et al. "The Angry Brain: Neural Correlates of Anger, Angry Rumination, and Aggressive Personality" (2009) 21 *Journal of Cognitive Neuroscience*, 734–744. It is unclear what sort of information discovering neural correlates really gives us about the nature and content of commonsense mental states. But most would agree that this sort of evidence indicates it is likely such mental states have some sort of physical instantiation in the brain.

the hard sciences: (1) this may be true, but a close examination of the progress made in brain science also shows slow progress toward an understanding of human cognitive capacities couched in commonsense psychological terms and (2) such a synthesis itself would not prove commonsense psychological concepts to be false. Instead, it would seem to show that commonsense concepts are real in that they reliably refer to the brain states we already know are causally linked to human behavior. To show commonsense concepts are radically false, eliminativists would be required to show that such concepts are in fundamental conflict with the way the hard sciences describe the world. But again, progress in brain science does not indicate this. Instead, progress is being made to identify the referents of mental states in the brain.

This is not to say that we should expect commonsense psychology to be a perfect, or even very accurate descriptor of the brain states that cause human behavior. As I will discuss in more detail later, commonsense concepts are instead likely to roughly refer to brain processes. And it ought not surprise us that commonsense psychology fails to adequately explain certain psychological phenomena, such as the way in which memories are made and stored. The apparent function of mind-reading and commonsense psychology is (at least in part) to understand and predict human behavior, which would require identification of mental states as brain states that have enough of a discrete identity to have reliable causal effects. This, in turn, would require recognition that persons can and do indeed form memories. However, it may not be necessary for the folk to understand how the cognitive processes vital to memory creation and storage operate. But it hardly follows from the fact that a theory is incomplete that the theoretical claims that are made are false.

To put it plainly, it seems the Churchlands' call to eliminate commonsense psychology as a theory of human behavior was imprudent. And this ought to be a relief to legal scholars, responsibility theorists, and anyone else who supports the project of holding persons responsible. This is because, as stated earlier, if commonsense mental states aren't real – in the same way that witches aren't real – then we are unjustified in finding persons guilty of a crime. The courts, in assessing criminal guilt, would be seeking fictional/nonexistent entities that (as such) cannot possibly have any impact on human behavior. If commonsense psychological concepts are eliminated, then the mental states the court is looking for in order to attribute criminal responsibility do not exist.

Elminativism about Free Will

Most philosophers now agree that Churchland-style eliminativism was ill-conceived. However, a subtler brand of eliminativism relevant to criminal responsibility has recently emerged: eliminativism about free will.[22] These eliminativists hold that free will doesn't exist, and free will is required for criminal responsibility; therefore, no one is ever criminally responsible. The position is different from that of the Churchlands, in that the new eliminativists do not advocate for elimination of commonsense mental states (in favor of some other, scientific understanding of the causes of human behavior). Instead they argue that our commonsense or folk responsibility practices, which underpin the criminal law, assume a special sort of causal power on the part of agents – a sort of agential control over their decisions and actions – that they do not actually have. In this case, argue the new eliminativists, it isn't fair or just to hold persons responsible for harm caused by such actions, or subject them to harmful punishment, because they aren't really the cause of such harm.

This is an interesting argument from the perspective of the mind body relationship, for it seems that the new eliminativists hold that mental states are indeed real things in the world.[23] However, our mental states do not have causal powers such that we can blame and hold people responsible for their actions. The folk notion of mental causation is radically false and ought to be eliminated as an explanation of human behavior. Once this is done, responsibility, including criminal responsibility, disappears.

Some have argued that the new eliminativist position is supported by experimental research into ordinary perceptions of free will. Certain studies seem to indicate that the ordinary folk don't tend to believe determinism is true, and when they are forced to consider the possibility that the universe is deterministic, the folk are less likely to apply praise and blame.[24] When presented with an explicit account of a determinate

[22] See Pereboom, n.13; Caruso, n.13; Caruso, G. *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will* (Lanham, MD: Lexington Books, 2012); Caruso, G. "Free Will Eliminativism: Reference, Error, and Phenomenology" (2015) 172 *Philosophical Studies* 10, 2823–2833; Pereboom, D. *Living without Free Will* (New York: Cambridge University Press, 2001).

[23] See Caruso, ibid,-2012 and Pereboom, n.13.

[24] Nichols, S. and J. Knobe "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions." (2007) 41 *Nous* (4), 663–685.

universe, one study showed that persons were hesitant to hold persons within that universe responsible.[25] However, other studies seem to have found that persons are still willing to hold agents responsible, even if those agents' actions appear to be determined.[26] Murray and Nahmias attempt to explain the conflicting data as confusion on the part of the folk regarding what determinism really means. This confusion, they say, fueled incompatibilist intuitions in cases where the study subjects were hesitant to attribute responsibility. If subjects are led to believe that determinism means that a person's conscious mental states are bypassed as a cause of their behavior, then responsibility intuitions are undermined. However, if subjects feel that an action is related to or caused by a person's conscious mental states, subjects seem willing to attribute responsibility, even if determinism is true. Because determinism does not entail bypassing the conscious self as a causal agent, Murray and Nahmias claim folk intuitions with regard to responsibility may remain intact even if determinism is true.[27]

It seems likely that folk intuitions about free will are mixed: that they are partly incompatibilist (hold determinism is incompatible with free will), and partly compatibilist (hold determinism is compatible with free will). Thus, if determinism turns out to be true, folk notions of free will may be partly true and partly false.[28] Nahmias and Murray's arguments, however, indicate that the folk notion of free will is very likely to be true enough to be preserved, despite determinism. A fairly liberal theory of reference would certainly seem to allow revision of the folk concept of free will such that it could be construed to refer to real brain processes and events corresponding to the conscious mental states that play a causal role in behavior. In this way the folk concept of free will might be more similar to the concept of *whale* – which had to be revised because humans initially assumed they were fish – than the concept

[25] Nichols, S. and A. Roskies "Bringing Moral Responsibility Down to Earth" (2008) 105 *Journal of Philosophy* (7), 371–388.

[26] Nahmias, E. and D. Murray "Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions" in J. Aguilar, A. Buckareff, and K. Frankish eds., *New Waves in Philosophy of Action* (London: Palgrave-Macmillan, 2010) 189–215 and Murray, D. and E. Nahmias "Explaining Away Incompatibilist Intuitions." (2014) 88 *Philosophy and Phenomenological Research* (2), 434–467.

[27] Murray and Nahmias, ibid.

[28] See Vargas, M. *Building Better Beings: A Theory of Moral Responsibility*, (Oxford: Oxford University Press, 2013).

witch, which had to essentially be eliminated because it turned out there were no women with magic powers.

Gregg Caruso and Derk Pereboom, however, have argued the opposite, claiming that the folk concept of free will is more like the concept *witch* than the concept *whale*. Two arguments regarding the radical falsity of folk notions of free will seem to be present in their work. First, Caruso claims free will historically refers to the *phenomenological feeling* of free will, which is essentially libertarian and incompatibilist in nature.[29] That is, the feeling of making a free decision indicates that our decisions are not determined (or subject to indeterminate forces). I disagree, however, that this feeling of freedom is central to the folk notion of free will, and I think my view is supported by the relatively small role the phenomenal feeling of freedom plays in responsibility assessments, a human practice that robustly engages folk notions of free will. The criminal law, for example, does seem to engage folk notions of free action, but in what sense? A defendant is not criminally responsible if the act was subject to duress, hypnosis, very low IQ, sleepwalking, or acts committed under the influence of a serious mental illness, and these are all circumstances where we might consider an agent's freedom to act impaired. But in each of these cases the defendant's cognitive system is either deviant when compared to a normal adult, or the defendant is constrained by very abnormal external circumstances. Only in the first case of duress does excuse correlate with a lack of feeling free. (I assume a hypnotized person and a sleeping person either feel free in their actions or feel nothing – I do not think they feel unfree. And I am quite sure some persons suffering from serious mental illness feel free.) Others who we excuse from full criminal culpability, including young children and those with low IQs, also feel quite free as they act, but their cognitive incapacities relative to a normal adult result in lesser culpability under the law. Further, many of those determined to be fully criminally responsible sometimes do not feel free: we often hold persons responsible for acts committed under serious emotional stress or under the pressure of a very weighty decision or very difficult circumstances, even though in such cases the agent may feel unfree.

In a later coauthored paper, Pereboom and Caruso offer a second argument against the folk concept of free will by claiming it refers to a type of agential control that persons do not have regardless of whether

[29] Caruso, 2015.

the universe turns out to be largely deterministic or somewhat indeterministic. They argue both compatibilism and libertarianism about free will are false because neither provides an account of agential control sufficient to vindicate the truth of free will. "Against the view that free will is compatible with the causal determination of our actions by natural factors beyond our control, we argue that there is no relevant difference between this prospect and our actions being causally determined by manipulators. Against event causal libertarianism, we advance the disappearing agent objection, according to which agents are left unable to settle whether a decision occurs and hence cannot have the control required for moral responsibility."[30]

I think Pereboom and Caruso are right to claim the folk notion of free will refers at least in part to agential control over action. Again, however, it seems that the folk concept of free action may refer to the causal efficacy of conscious mental states as the locus of this control. Agents feel very strongly aligned with their conscious values, plans, and decisions, and such values, plans, and decisions are indeed causally related to much of human behavior.[31] And, the folk notion of freedom present in the criminal law seems clearly aimed at these types of causally efficacious mental states (classified as "purposely" and "knowingly," etc.). As mentioned earlier, cases of legal excuse tend to be cases where there is something wrong with the way in which these conscious states are formed and executed. For example, in cases of mental illness, addiction, and sleepwalking the states causing action are "abnormal" because they are not subject to the typical conscious cognitive review process, and in the case of young children and those with very low IQs, this system of review is underdeveloped.[32]

[30] See Pereboom, 2001; Caruso, 2012; and Pereboom and Caruso, 2018.

[31] This is not to deny that there are many unconscious mental factors that influence behavior. However, there are theories – my colleagues and I have a scientifically based version, and Angela Smith has a purely philosophical one – that the possibility of conscious rational endorsement is enough to assign responsibility to the mental cause of an act. See Smith, A. "Attitudes, Tracing, and Control" (2015) 32 *Journal of Applied Philosophy* (2), 115–132.

[32] See Fagan, T., W. Hirstein, and K. Sifferd "Innocent Minds: Child Soldiers, Executive Functions, and Culpability" (2016) 16 *International Criminal Law Review* (2), 258–286; Sifferd, K. (2012). "Translating Scientific Evidence into the Language of the Folk: Executive Function as Capacity-Responsibility" in N. Vincent ed., *Legal Responsibility and Neuroscience* (Oxford: Oxford University Press, 2012); Sifferd, K., W. Hirstein, and T. Fagan "Legal Insanity and Executive Function" in M. White ed., *The Insanity Defense: Multidisciplinary Views on Its History, Trends, and Controversies* (Santa Barbara, CA: Praeger, 2017), 215–242; Hirstein, W. and K. Sifferd. "The Legal Self: Executive Processes and Legal Theory" (2011) 20 *Consciousness and Cognition* (1), 156–171.

Caruso and Pereboom say that conscious causing isn't sufficient to ground agential control because it is no different than being causally influenced by intentional manipulators. It is well established that our conscious mental states are influenced by a whole host of factors outside the agent's control,[33] and this includes the possibility of malevolent manipulators. But it isn't clear that the fact that agents are influenced by outside forces is relevant to our assessment of the reference of the folk notion of free action. Caruso and Pereboom may feel the folk fail to distinguish between "normal" outside influences, like parenting and schooling, and "abnormal" influences, such as the manipulations of evil neuroscientists. But the folk concept of free action does not need to operate in a manner considered fair or consistent by philosophers for it to truly refer. The folk notion of "tasty" is hardly consistent (in the sense of treating like cases alike) with result to outcome: my concept of "tasty" successfully refers to strawberries, pixie sticks, and tequila, and one can hardly identify anything these tasty items have in common except the fact that I find them tasty. But this doesn't mean the concept of "tasty" ought to be eliminated: it truly and reliably refers to things in the world and refers accurately enough for us to use the concept to understand and predict our own and others' behavior. Philosophical arguments that a folk concept like "tasty" or "free action" fails to treat like cases alike may indicate a need to revise our understanding of the referents *in order to achieve certain social or ethical goals*. However, such an argument cannot provide grounds to eliminate either concept.

I believe that our responsibility practices are fair in large part, and that folk concepts are well equipped to treat like cases alike with regard to human action, but I need not argue for this here. All I need to argue to defend against an eliminative challenge is that the folk concept of free action does indeed reliably refer to a category of human behavior, and there is good proof of this in the criminal law. The folk concept of free action appears to refer – at least in part – to action causally related to the conscious mental states of a normal adult's cognitive system.

[33] See, for example, the enormous amount of research that has been done on the way in which "situational factors" impact persons' decisions.

Many legal scholars have similarly argued that responsibility practices of the criminal law do not demand libertarian or other false notions of free will or agential control.[34] As Stephen Morse claims:

> [T]he folk-psychological model of the person that is central to our explanations of human behavior and to responsibility doctrines and practices is not challenged by determinism in general or by neurodeterminism in particular. Criminal responsibility doctrines and practices are fully compatible with the truth of determinism (or causal closure). Until science conclusively demonstrates that human beings are not responsive to and cannot be guided by reasons and that mental states do not play even a partial causal role in explaining behavior, the folk-psychological model of responsibility will endure as fully justified.[35]

Morse agrees that legal responsibility doesn't rest upon whether a defendant felt free, or on a special kind of agential control. Instead, the law holds persons responsible who cause harm via their own conscious mental states. Folk concepts underpinning criminal responsibility also seem to track what Morse calls "minimal rationality"[36] and what philosophers call "reasons-responsiveness."[37] Some have argued that consciousness is what enables this rational capacity in human beings. When a mental state becomes conscious, it can be subject to review by the agent's long-term values and desire.[38] Morse argues that the criminal law

[34] See: Morse, S. "Criminal Responsibility and the Disappearing Person" (2007) 28 *Cordozo Law Review*, 25–45; Morse, S. "Determinism and the Death of Folk Psychology: Two Challenges to Responsibility from Neuroscience" (2008) 9 *Minnesota Journal of Law, Science and Technology* (1), 1–36; and Morse, S. "Neuroscience, Free Will, and Responsibility" in W. Glannon ed., *Free Will and the Brain: Neuroscientific, Philosophical, and Legal Perspectives* (Cambridge: Cambridge University Press, 2015).

[35] See Morse, 2015 at 253.

[36] Morse says, "The legal view of the person does not hold that people must always reason or consistently behave rationally according to some preordained, normative notion of rationality. Rather, the law's view is that people are capable of acting for reasons and are capable of minimal rationality according to predominantly conventional, socially constructed standards. The type of rationality the law requires is the ordinary person's common-sense view of rationality, not the technical notion that might be acceptable within the disciplines of economics, philosophy, psychology, computer science, and the like." Morse, ibid. at 256.

[37] Fischer, J. M. and M. Ravizza *Responsibility and Control: A Theory of Moral Responsibility* (Cambridge: Cambridge University Press, 1999).

[38] Neil Levy argues that for an agent to be held responsible for an action it must be causally related to conscious mental states; Levy, N. *Consciousness and Moral Responsbility* (Oxford: Oxford University Press, 2014). However, my colleagues and I argue that culpable action must be causally related to states that are *available* to conscious executive review – they needn't necessarily be conscious themselves. See Hirstein, W. and K. Sifferd, n. 32, 2011; Sifferd, n. 32, 2012; and Fagan, Hirstein, and Sifferd, n. 32, 2016.

[39] See Levy, ibid.

seems to reflect the standard compatibilist position: that some subset of determined actions/actions influenced by factors outside the agent can be assigned to the agent such that she may be held responsible for them. At least from the perspective of legal responsibility practices, the free will that matters is the capacity to act in accordance with reasons, including moral and legal reasons.

It seems, at least as an argument for eliminativism about mental causation, the Caruso and Pereboom position fails if the folk consider free action that action caused by conscious mental states (or mental states available to consciousness) such that such states are related to a minimally rational, or reasons-responsive, cognitive system.[39] Even if some of the folk intuitions refer to libertarian free will and thus are incorrect, it seems the core aspects of the concept are probably correct – especially, the aspects that seem especially important to the criminal law – and thus the concept of free will need not be eliminated. A whale may not be a fish, but it is still a very large gray animal that lives in the ocean.

Both forms of eliminativism discussed earlier – about mental states and agential causation – offer a theory of the mind–body relationship that poses a threat to mental causation. If mental states aren't real, they can't cause behavior. Further, if agents don't cause behavior in the way the folk think they do, then mental causation is false, and persons cannot be held responsible for their acts. To put it another way, if, after exploring the eliminativist arguments, we were convinced that either version of eliminativism was true, it would be hopeless to seek a theory of the mind–body relationship that preserves mental causation. However, it seems that both eliminative threats can be avoided: neither the folk concepts of mental states, nor the folk concept of free will, are radically false. Thus, such concepts can be preserved to ground responsibility attributions made in the criminal law. Our task now is to seek a positive theory of the mind–body relationship that preserves the mental as a real thing in the world with causal powers. This theory should help us understand how the mental possesses causal powers via its physical realizers.

Non-Reductive Physicalism/Emergence

Before discussing non-reductive physicalism, let me say a few words about why I have dismissed ontological dualism as a possible theory of the mind–body relationship that can preserve mental causation. Ontological dualism – the idea that the mental is of fundamentally different stuff than the physical – poses well-documented problems for mental causation.[40] If mental states occupy another ontological category than the physical body, it becomes very difficult to explain how the immaterial mind has physical causal effects.[41] As Princess Elizabeth of Bohemia commented to Rene Descartes in a letter in 1643, there seems to be no way to conceive of how an immaterial soul or immaterial mental states can have any effect on material substances, including the material substance of the human body. "For the determination of movement seems always to come about from the moving body's being propelled – to depend on the kind of impulse it gets from what sets it in motion, or again, on the nature and shape of this latter thing's surface. Now the first two conditions involve contact, and the third involves that the impelling thing has extension; but you utterly exclude extension from your notion of soul, and contact seems to me incompatible with a thing's being immaterial."[42]

Philosopher Colin McGinn has said the following about dualism about the mind:

[40] See Davidson, D. "Actions, Reasons, Causes" (1963) 60 *Journal of Philosophy* 685–700. See also Mele, A. *Springs of Action* (New York: Oxford University Press, 1992).

[41] See Kim, J. "The Myth of Nonreductive Materialism" in P. K. Moser and J. D. Trout eds., *Contemporary Materialism* (New York: Routledge, 1995), 134–149; Sifferd, K. "What Does It Mean to Be a Mechanism? Stephen Morse, Non-Reductivism, and Mental Causation" (2014) *Criminal Law and Philosophy*, 1–17. Theories that deny an ontological reduction are thought by many to render the mental epiphenomenal. Due to the causal closure of the physical realm, physical states seem to be the obvious winners if the mental is forced to compete with regard to the causation of behavior. See Papineau, D. *Thinking about Consciousness* (Oxford: Oxford University Press, 2002). If commonsense mental entities or events cannot be understood or explained in terms of their physical underliers, then we cannot reliably attribute causal powers to them. The claimed epistemological gap threatens assignment of causal powers to the mental because such an attribution of causal powers seems excluded by the assignment of causal powers to the physical underliers. See Kim, J. *Physicalism, or Something Near Enough* (Princeton, NJ: Princeton University Press, 2005).

[42] Quoted in: Anscombe, E. and P. Geach *Descartes: Philosophical Writings* (Indianapolis, IN: Bobbs-Merrill Company, 1954).

> Dualism proposes to give the mind its ontological due, but the problem is that it has difficulties organizing a rendezvous between the two spheres: how does the mind affect the brain and the brain the mind? Whence the systematic correlation and interaction? . . . Dualism makes the mind too separate, thereby precluding intelligible interaction and dependence.[43]

As McGinn indicates, no sound theory has emerged explaining how an immaterial substance could create physical causal effects, and thus provide a plausible account of mental causation. However, McGinn thinks dualism gives the mind "its ontological due" because it holds what McGinn, and some other philosophers and legal theorists, feel must be true about the mind – that it cannot *just be* physical states or processes; that instead, there must be some important, special aspect of the conscious human mind that is separate from, and thus cannot be reduced to, its underlying physical realizers.[44] Even among philosophers committed to physicalism, this notion of the fundamental separateness of the mental from the physical seems salient, and has fueled non-reductive positions about the mind that have gained popularity in the past few decades. To my mind, the appeal of such theories is that they appear to allow theorists to have their cake and to eat it too: the mind remains "special" such that it cannot be reduced to the physical realm, but it doesn't occupy a new (nonphysical) ontological category. Non-reductive physicalism thus seems to represent a sort of metaphysical middle position.

The question posed here is what sort of theory of the mind–body relationship will preserve the mental causation required for criminal verdicts. My tactic in this section will be to closely examine two non-reductive theories that posit a *close* relationship between the mental and the physical: Donald Davidson's non-reductive physicalism and John Searle's biological naturalism. Searle argues that conscious states cannot be ontologically reduced to physical states, and Davidson argues that mental states resist an epistemological reduction. If these two non-reductive theories fail to preserve mental causation, despite the close relationship they posit between the mental and physical,

[43] McGinn, C. "All Machine and No Ghost?" *New Statesman*, 2012.

[44] This sentiment seems to fuel Stephen Morse's acceptance of non-reductive physicalism. He holds that mental states are not reducible because they "are realized in the brain – the mind-brain – but not at the level of neurons," Morse, n. 3, 2008: at 33.

I don't think we should hold out much hope that other non-reductive theories will fare any better.

Non-reductive physicalism is the position that although everything that exists is physical, and thus the mental is physical, some aspect of the mental cannot be reduced to the physical. Thus, to explore non-reductive theories about the mental requires exploration of the notion of reduction. Generally, reductions may be ontological or epistemological. Ontological reductions can be made between entities, objects, events, processes, and properties, and epistemological reductions are made between concepts, theories, models, and frameworks.[45] An ontological reduction is made between real-world entities, and a strong claim about ontological reduction would indicate that everything in the world can be said to be nothing more than or determined by some fundamental constituent of reality (e.g., atoms or the entities posited by quantum physics). An epistemological, or theoretical reduction might entail fully explaining or understanding higher-order mental states or events in terms of a science of lower-level brain states or events.

A non-reducible entity, property, theory, or framework is often said to be *emergent*, which is just another way of saying that it cannot be reduced. Thus both dualist and non-reductive positions on the mind–body relationship posit that some aspect of the mental is emergent and cannot be reduced to the brain. Davidson's brand of non-reductive physicalism is best understood as acceptance of ontological reduction and denial of an epistemological/representational reduction.[46] Searle's non-reductionism, on the other hand, denies an ontological reduction, while at the same time, trying to maintain its status as a physicalist theory. As we shall see, this isn't easy. One might wonder: if some aspect of mind occupies a separate ontological category from that of the physical brain or body, what sort of (physical) thing could it be?

Davidson's Non-Reductionism

Donald Davidson's non-reductionism is perhaps the most influential version of the theory. Davidson began to argue for a non-reductive physicalist position in the late 1960s and 1970s, partly in reaction to

[45] See Van Gulick, R. "Reduction, Emergence and Other Recent Options on the Mind/Body Problem: A Philosophic Overview" (2001) *Journal of Consciousness Studies* 8 (9–10), 1–34 at 3.4.

[46] Van Gulick, ibid.

the perceived failures of the type-identity reductive project, which aimed to solve the problem of mental causation by identifying types of mental entities with physical types, such as brain states.[47] Davidson argued the relationship between the mental and the physical is not that of identity (where a mental state *just is* some brain state or other) but instead argued for a relationship of weak supervenience. Specifically, Davidson argued that the mental supervenes on the physical, such that any change in a mental property has to be accompanied by some change in a physical property. However, the reverse does not hold: a change in physical properties does not necessarily entail a mental change.[48] This means that although each mental event has some physical identity or other, the same mental event could be instantiated as very different physical brain events: a mental state such as thirst is "token" identical to some physical state, but it may be identical with different physical states within and across persons.

The motivation for Davidson's position is "anomalous monism." Although all events occur within the single physical universe (the position of monism), physical and mental events seem to participate in very different types of laws. While scientific theories or frameworks seemed to posit entities governed by strict causal physical laws, the folk concepts and theories utilized when positing mental events do not seem so governed.[49] For example, I don't always go get a drink when I am thirsty – I might not get a drink if I were in the middle of teaching a class, or if I were running a race. Therefore, the mental and physical realms were deemed by Davidson to be incompatible insofar as there could not be strict psychophysical laws: explanations utilizing the ceteris paribus laws governing mental events could not be understood in terms of physical events, which appeal to the laws of physics. In other words, the laws of the special sciences (sciences focused on understanding human behavior) are so fundamentally different from those governing other things in the world that mental explanations of behavior cannot be reduced to (read: understood in terms of) physical explanations of behavior.[50]

Davidson's version of non-reductionism has been criticized on various grounds, but the theory is particularly worrying with regard to mental causation. The problem with Davidson's theory is that if the same mental state can supervene on different physical states, the mental

[47] See Smart, J. "Sensations and Brain Processes" (1959) *Philosophical Review* 66, 141–156.

[48] See Davidson, D. *Essays on Actions and Events* (Oxford: Clarendon Press, 1980).

[49] Davidson, ibid.

[50] Davidson, ibid.

cannot reliably substantiate causal properties because these properties are instantiated within the physical realm. As Jaegwon Kim has noted, Davidson's position means that the very same network of physical causal relations would obtain even if one were to redistribute mental properties over it randomly; on Davidson's theory "... you would not disturb a single causal relation if you ... removed mentality entirely from the world."[51]

Davidson acknowledges that the physical realm operates in accordance with strict causal laws, and that physical states are causally related to behavior. So, when I get up to get a drink, there is a physical causal explanation for my action that adheres to strict laws. There is, of course, also a mental explanation: I was thirsty; I knew there was a drinking fountain in the hallway; thus, I decided to go get a drink of water. But at the heart of Davidson's theory is the claim that the mental state of being thirsty can be instantiated in different ways in the brain from one tokening to the next. So my thirst and belief that there is a drinking fountain in the hallway might be instantiated in my, say, in my prefrontal cortex in one moment, but the next time I am thirsty, this thirst supervenes on my amygdala; and and my belief that there is a drinking fountain in the hallway supervenes on my visual cortex.

However, if this were the case (1) it isn't at all clear how the mental state of thirst can have reliable behavioral effects, such that if you know I am thirsty you might predict I was going to get a drink, and (2) it isn't clear how the mental explanation is a causal explanation of my behavior, given that the physical explanation is most certainly causal, and the mental explanation only attaches to the physical explanation in an unprincipled way (or, to put it plainly, willy-nilly). In sum: the physical causal nature of the brain is indeed playing a primary causal role in our behavior, and if we cannot reliably link mental states to these causal processes, the mental does not act as a cause of our behavior. Thus, on Davidson's theory the mental ends up looking as though it is epiphenomenal (has no causal power). But from the perspective of the law, when judges or juries look for the mental states that caused criminal harm, they are looking for a causal relationship between mental event and criminal harm that doesn't reliably exist. For this reason, Davidson's theory does not posit a relationship between the mind/brain that can support criminal verdicts.

[51] See Kim, n. 41, 1995 at 136.

Searle's Non-Reductionism

I have discussed Searle's non-reductionism in detail elsewhere.[52] Here I will attempt to give a short summary of some of my prior arguments. Searle's non-reductionism is interesting with regard this chapter's project because one of Searle's expressed aims in articulating this theory is to preserve mental causation. Searle's theory, termed "biological naturalism," consists of four theses:[53]

1. Consciousness cannot be reduced to its neurobiological basis because "such a third-person reduction would leave out the first-person ontology of consciousness."
2. Conscious states, however, are causally reducible to neurobiological processes – they are not something "above and beyond" such processes.
3. Portions of the brain system composed of neurons are conscious, although individual neurons are not.
4. Conscious states function causally – they have physical effects.

The causal efficacy of mental states (thesis four) relies upon their causal reducibility (thesis two). Thus, Searle says conscious mental states have physical effects by virtue of the place they occupy within a physical causal system: the causal properties of a conscious state are identical with the causal properties of the physical state that realizes it.[54]

Searle maintains a non-reductive position despite his claim that mental entities are causally reducible by arguing that the *conscious phenomenal* properties of a mental state are emergent.[55] The very point of the concept of consciousness is to capture the first-person, subjective experience, claims Searle, and this subjective phenomenal feel of a conscious state is not constituted by, or identical to, underlying brain states or processes. Instead, phenomenal feel is an ontologically emergent property of biochemical brains like ours. Thus, while Davidson admits an ontological reduction and denies an epistemological one, Searle denies an ontological one, at least with regard to phenomenal feel.[56]

[52] See Sifferd, 2014, n. 41.
[53] Searle J. *Mind: A Brief Introduction* (New York: Oxford University Press, 2004).
[54] Ibid. at 118. [55] Ibid.
[56] Ibid. Again, an ontological reduction involves a claim that an object or event of a certain type can be shown to be nothing more than or determined by objects or events of another type, in the way that a rock can be shown to be nothing more than atoms arranged in a certain way.

Searle claims that consciousness is a "causally emergent property" of systems of neurons, whose existence can be explained by the causal interactions between elements of the brain at the microlevel, but can only be accessed by the person who owns the brain producing consciousness.[57] But this argument seems odd given that Searle denies property dualism – that is, he denies that there are nonphysical properties. This means that the properties of consciousness must be physical spatial properties of the world with causal effects. But in this case, it seems exceedingly mysterious how certain physical and spatial properties of the brain that have physical causal effects can be accessible only to consciousness itself.[58] In other words, Searle's position is ontologically very odd: it is unclear how such phenomenal feels, which he admits are physical causal properties, are properties that are necessarily invisible to science.

Of course, many would agree with Searle that empirical examination of a brain state (using current scientific means) will leave out information of how this brain state feels to the holder of the state. But this is an argument against epistemological, not ontological, reduction. And as Van Gulick notes, one oughtn't to infer that mental properties are ontologically emergent even if it turns out to be the case that we cannot representationally reduce our mental concepts or theories to physical ones.[59] To do so would be to assume an epistemological divide is enough to prove an ontological one. And this is what Searle seems to do: he resists an ontological reduction based on his notion of a subjective/objective divide. This move seems especially suspect given that Searle causally identifies mental states with physical states. One important way to divvy up ontological entities is via their causal powers. Thus, one might argue for ontological emergence by claiming that wholes or systems have causal powers that are emergent from the powers of their parts, such that the system-level causal powers at the macromental level were not determined by the powers of their parts at the microbrain level.[60] With regard to the mind–body relationship, it could be claimed that macrolevel entities, such as mental states, somehow exhibit causal powers not exhibited or determined by their underlying lower-level neuronal states, indicating a real ontological divide.[61] But Searle argues that the causal properties of mental states are identical to the causal

[57] See Corcoran, K. J. "The Trouble with Searle's Biological Naturalism" (2001) *Erkentnis* 55 (3).
[58] Ibid. [59] Van Gulick, n. 45 at 28. [60] Ibid. at 17. [61] Ibid.

properties of brain states. If true, such identical causal properties might be interpreted to provide substantive support for an ontological reduction.

What does Searle's theory mean for mental causation? Mental causation requires that the intentional content of mental states be a causal explanation of my behavior. Searle's claim that the causal properties of mental states are identical to their underlying brain states has an interesting result: any aspect of a mental state that carries causal properties – presumably, their content – carries them via underlying microlevel brain states. The flip side of this is that the aspect of the mental Searle says isn't reducible – phenomenal consciousness – can do no causal work in addition to the mental content as causally realized at the microlevel. That is, Searle's theory indicates that conscious mental states qua their conscious properties – those conscious properties only accessible to the state-holder and not to science – can do no causal work because they are non-reducible. So on Searle's theory, conscious mental states do causal work only insofar as they are related to causal mental content, which he says can be ontologically reduced (at least, Searle has provided no argument that they can't be ontologically reduced). Thus conscious properties, as Searle paints them, are epiphenomenal: the conscious mental state of thirst does not cause me to get a drink – some underlying physical realizer of the causal mental content does. This means that the conscious mental states a judge or jury are looking for when trying to determine if an offender is guilty of a crime are also epiphenomenal: that is, not causally related to the criminal harm the offender was arrested for. In the end, Searle's theory, like Davidson's, does not preserve mental causation and, thus, cannot support criminal verdicts.

The Non-Reductivist Project and Mental Causation

I argued above that two different attempts at a non-reductionist theory – Davidson's, which denies an epistemological reduction, and Searle's, which attempts to deny an ontological one – fail to preserve mental causation. As Tim Crane has noted, in general, non-reductive theories seem to create problems for mental causation because they force higher-level mental properties to compete for causal power with their lower-level physical realizers.[62] Thus, the problem non-reductive physicalism creates

[62] Crane, T. "Mental Causation" in L. Nadel ed., *Encyclopedia of Cognitive Science* (New York: Wiley, 1995).

is the one physicalism was intended to solve. Physicalism was initially posited as a *solution* to the problem of mental causation. As Smart noted in his famous 1959 paper, mental states seem to cause physical effects, but physical effects have complete physical causes. Therefore, Smart claimed, physicalism must be true – to have causal powers, mental states must be simply recognized as physical states.[63]

As we saw earlier, the non-reductivist argues that although mental concepts or entities have physical causal effects via their relationship with the physical entities, these causes and effects cannot be understood in terms of those physical entities via a scientific theory or scientific concepts. As Van Gulick says, "The challenge [for non-reductivists] has been to show that one can 'eat one's pluralist cake' while still remaining robustly physicalist at the ontological level." Specifically, non-reductivists must show that explanations exist at the level of folk psychology – with regard to causal explanations of intentional behavior – that cannot be understood or accessed by use of physical theory.[64] As I noted in an earlier paper, this seems problematic: if the gap posited by the non-reductionist between the mental and physical is ontological, then looking for a particular mental state that caused an action is a fool's errand because there is nothing to be found fitting that description.[65] If, however, the gap is epistemological, then looking for a mental state that caused an action is still a fool's errand because although there might be such a mental state, we are (somehow) necessarily blind to it. In neither case can mental causation, and thus responsibility, be preserved.[66]

Non-Eliminative Reductive Physicalism

Legal scholars Dennis Patterson and Michael Pardo say this about the possibility of reduction of the mental state terms used in law:

> Like neuro-reductionists generally, neurolegalists aspire to reduce the explanation of all human behavior to the level of brain processes. Believing as they do that "the mind is the brain," neurolegalists have attempted to account for mental capacities, abilities, and processes solely at the level of cortical function. As an explanatory account of the nature of human agency, this reductionism aspires to nothing less than the replacement of our talk of beliefs, desires, and intentions with the language of neuroscience.[67]

[63] See Smart, n. 59 and Papineau, n. 17. [64] Van Gulick, n. 45 at 22. [65] Sifferd, n. 41.
[66] Ibid.

Pardo and Patterson claim that a reductionist account is going to eliminate and replace our commonsense mental state concepts. But I will argue here the exact opposite is true. A non-eliminative reductionist need not replace and eliminate the folk concepts that underpin responsibility. Instead, reductionist accounts can *vindicate* these concepts and preserve the full-bodied mental causation necessary for responsibility assessments. If a folk psychological mental state is identified with the physical, or the physical is deemed constitutive of the mental, there is no gap we need to overcome to assign the mental causal power. Instead, under non-eliminative reductionism, mental states carry causal properties *just as* physical entities.

Non-eliminative reductionism rests on the notion that a mental state at time *t* is some physical (probably, brain) state. Thus, an agent's desire for a drink of water just is a physical state or process within a particular time slice – no more, no less. This understanding of reduction indicates that all reductions are necessarily non-eliminative. Non-eliminative reductionists do not worry that the mental will disappear or be eliminated via a reduction, just as scientists did not worry that the concept of "water" needed to be eliminated once it was discovered to be H_2O. Even a "bumpy" or "lumpy" reduction does not require a folk concept to be eliminated: only a failed reduction results in elimination. Even assuming physicalism, and thus an ontological reduction of a mental state to some physical state, an epistemic reduction of one theory or framework to another is often fairly rough. As Michael Moore notes, "Even temperature's reduction to mean kinetic energy, for example, was said to have a little lumpiness in it, because of the need to stipulate a randomness of direction in molecular motion in order to reach the identification needed. Lumpier still, it was thought, was the 'mass' in Newton's theory in relation to the mass of which post-relativistic physics speaks; one has to work a bit on what mass means in Newtonian mechanics to get that to be but a special case of Einstein's theories" (see Michael Moore, this volume, Chapter 2).

Of course, if the epistemic reduction proves to be too lumpy – if there is no path to map one set of theoretical entities onto the other – a

[67] Pardo M. and D. Patterson *Minds, Brains, and Law* (Oxford: Oxford University Press, 2013), 28–29.

reduction must be rejected. For example, if marine biology discovered that whales were actually millions of tiny fish stuck together that only appeared to move and live as one organism, it would seem that our concept of whale might be so radically false that we would be required to abandon it. We would say things like: "What we thought were whales turned out to be millions of minion fish." But a bit of lumpiness is to be tolerated, even expected: the concept *whale* was not eliminated when we discovered they aren't fish; instead, we revised our concept of whale to refer to large gray *mammal* that lives in the ocean. In the first case, the existence of whales was undermined by advancing science; in the second case, our folk concept of whale was revised – both vindicated and made more accurate – by advancing science.

The difficulty of a theoretical reduction of mental states to physical states may be exaggerated because for millions of years, subjective experiences of mental states failed to provide evidence of their physical reality. Humans discovered that brains were the locus of mental states long after they knew what mental states felt like, and long after we used mental states to understand and predict our own behavior. However, as Saul Kripke has noted, identity claims need not be known a priori. They can be known a posteriori, or with experience.[68] Human beings did not understand that water was H_20 until we had access to certain scientific tools; similarly, we only came to understand that thirst is some brain state or other via use of certain scientific tools (namely, the tools of contemporary cognitive science and neuroscience). The scientific tools now available to us have allowed us to closely examine the nature of conscious mental states, and it seems the best evidence indicates that the immediate subjective experience of a mental state may be *just what it is like* to have or undergo a particular physical state.[69]

The position of non-eliminative reductionism holds that the causal relationship between folk psychological mental states (such as thirst) and actions (such as getting a drink) can be understood on multiple theoretical levels: for example, at a macrobehavioral level of folk explanations, and at a microlevel, such as the level of functional neuro-connectivity or neurons. Each level may portray a true causal explanation regarding my getting a drink.[70] Understanding the relationship between levels is

[68] Kripke S., *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1980).

[69] Certainly, our current best science has not provided evidence that the subjective feeling of a mental state is anything more than what it is like to undergo that mental state.

[70] Van Gulick, 2001, n. 41.

difficult, complicated work, but several philosophers have already attempted such projects.[71] The non-eliminative reductionist claims this is work that can be done, precisely because there is no unbridgeable gap between mental states couched in folk terms and their physical realizations. The position of non-eliminative reductionism holds that an epistemological reduction of mental states is possible, even though mental states couched in folk terms and brain states enter into fairly distinct systems of laws. Again, it seems very likely this reduction will be bumpy, and that folk concepts may need to be revised, but core folk psychological concepts will not need to be eliminated.

Some have worried that commonsense mental state terms pick out such messy physical states that there may not be relevant physical commonalities that would allow an identity relation. Obviously, a reductionist theory must admit the difficulty of understanding folk psychological concepts in scientific terms, which would be at best an extremely complex process, especially given the breadth and depth of our folk psychological concepts and explanations. One of the reasons some philosophers turned away from type identity theory in the past few decades (and toward non-reductionism) is because of this sort of worry, exemplified by Hillary Putnam's multiple-realization objection. Because pain seems to be realizable by very different physical systems (across different species and persons, and potentially in alien systems) it would seem that pain as a type cannot be identical with some particular type of physical event.[72] That is, it is very difficult to identify a physical commonality between human pain and octopus pain, such that "pain" has a physical identity.

However, it is unclear multiple-realization was a problem worth creating the even deeper problems of non-reductionism (namely, problems for mental causation).[73] There are various ways around the multiple-realization objection. One could argue, for example, that pain is a functional concept, but that local identifications must be made with whatever first-order physical properties or events that instantiate the functional concept within a particular system. In this case mental

[71] See Craver, C. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience* (Oxford: Oxford University Press, 2007); Kim, J. "Events as Property Exemplification" in *Supervenience and Mind: Selected Philosophical Essays* (Cambridge: Cambridge University Press, 1993); Papineau, n. 17, 1993; and Roskies, A. "Decision-Making and Self-Governing Systems" (2016) *Neuroethics*, 1–13.

[72] Putnam, Hilary *Representation and Reality* (Cambridge, MA: MIT Press, 1988).

[73] Sifferd, n. 41.

properties may be understood as second-order properties, the property of having some further physical properties the instantiation of which plays a certain functional role. Papineau, Lewis, and Kim[74] all use this sort of strategy, arguing that even if multiple realizability means broad type identity claims for psychological kinds such as "pain" aren't possible, local reductions are.[75] According to this view, a psychological kind might reduce to a local neuroscientific (or other scientific) kind. In this case the folk concept of "human pain" refers to a local disjunction of physical states, whereas "octopus pain" refers to a different local disjunction.[76]

Further, as I have argued at various points earlier, non-eliminative reductionists can argue that folk psychological concepts are flexible. The folk may not have realized their folk psychological concepts such as thirst picked out disjunctive local reductions that can be described on many different levels using different theoretical tools. Nonetheless, this appears to be the case, and this realization does not undermine the truth of folk concepts. Instead, these new facts explain why the concepts actually do work to predict and understand human behavior. To take an example from Eliot Sober: it seems true that smoking causes cancer due to some microconfiguration in cigarette smoke. It may also be true that different types of cigarettes have differently-realized cancer causing components at the microlevel. Even so, the macrolevel claim "smoking causes cancer" is true and explains each of the singular instances of cancer caused by cigarettes.[77] An ideally complete explanation of a singular occurrence would include the macrostory, the microstory, and an account of how they are connected. Sober notes that "the kernel of truth in the multiple realizability argument" is that higher-level descriptions may "abstract away" from the physical details that make for differences among the

[74] Papineau, n. 17, 1993, at 25; Kim, n. 71; and Lewis, D. "Reduction of Mind," in Samuel Guttenplan ed., *A Companion to Philosophy of Mind* (Oxford: Blackwell Publishers, 1994), 412–431.

[75] See Funkhouser, E. "Multiple Realizability" (2007) 2 *Philosophy Compass*, 303–315.

[76] Michael Moore also seems to hold this view. This is what he says about the multiple realizability challenge in Chapter 2 of this volume: "[S]urely it will not prove to be the case that some mental state *M* can be realized by some distinct brain state *B* that differs on each occasion that *M* exists. Surely, that is, *M* is realized by a finite disjunction of brain states, B_1, or B_2, or B_3. . . or *Bn*, where n is not infinite and probably not even that large. If so, then *M* will be type-identical to the finite disjunction of brain states, even if it is not type-identical to any one disjunct."

[77] Sober, E. "The Multiple Realizability Argument against Reductionism" (1999) *Philosophy of Science* 66(4), 542–564.

microrealizations that a given higher-level property possesses.[78] However, Sober claims, this does not make higher-level explanations "better" in any absolute sense. Different levels of description have different benefits.

An interesting result of epistemic reductionism is that it can offer the opportunity to pick from multiple levels of true theoretical understandings of human behavior depending on our task. One benefit of folk psychological explanations of behavior is that they are intuitively available to human beings and tied to other important human activities, such as responsibility assessments. As Adina Roskies notes, the multirealizability of folk mental state concepts can be understood to highlight the importance of commonsense psychological explanations, instead of undermine them. Physical states are more finely grained than the mental states they realize, and because folk concepts refer to a disjunction of physical states, the same psychological state can supervene on (somewhat) different physical states across time and persons. Because giving an account of mental causation in physical terms would require reference to a messy disjunction of physical states, it makes sense that human beings tend to use folk psychological terms when understanding and predicting behavior because folk explanations "capture much more succinctly the relevant counterfactuals."[79] To put it another way: explanations of human behavior referencing psychological states express the right level of generality to capture counterfactuals relevant to explanations of human behavior; explanations using brain states do not. This would seem to indicate that the criminal law is correct to use commonsense mental state terms as the primary means to attribute responsibility.

Once we accept both an ontological and a theoretical reduction, the causal properties of mental states become much easier to understand. If my thirst is a physical event or state in my brain, then it holds causal properties identical to this brain state. As some proponents of physicalism noted, human action seems to have a full physical/causal explanation at the microlevel.[80] Even if portions of this physical explanation are indeterministic, the causal chain resulting in an action can be understood

[78] Ibid. at 560.

[79] Roskies, A. "Don't Panic: Self Authorship without Obscure Metaphysics" (2012) 26 *Philosophical Perspectives*, 323–342.

[80] See Smart, n. 47; Papineau, n. 41, 2002; Papineau, D. "Causation Is Macroscopic but Not Irreducible" in S. Gibb, E. J. Lowe, and D. Ingthorsson eds., *Mental Causation and Ontology* (Oxford: Oxford University Press, 2013), 127–151.

as microcausal precursors to that action. Non-eliminative reductionism holds that some subset of those prior microlevel causes of an action – such as getting a drink – *just are my mental states (e.g., my thirst)*. One might imagine being able to look backward, bracket them off, and point at them in awe: "Look, there they are! My conscious mental states! Look at how they caused my behavior!" Or, we might imagine a different scenario where mental states sought after may not be found. For example, in the process of determining if a defendant is criminally responsible, the court might say: "Looking back at the time of the crime, it seems unlikely the defendant had the desire to kill his neighbor. There just no evidence that he possessed this mental state."

Earlier I argued two different versions of non-reductionism fail to preserve mental causation. Davidson's anomalous monism failed to reliably link mental states to the physical states that indubitably carry causal properties. Thus, on Davidson's theory the mental ends up looking as though it is epiphenomenal. Searle claimed that the causal properties of mental states are identical to their underlying brain states, but refused to reduce their conscious phenomenal properties, so conscious mental states qua their conscious properties could do no causal work. Non-eliminative reductionism bites the bullet non-reductivism tries to avoid by claiming that all aspects of mental states, including their causal properties and their phenomenal feel, are ontologically and epistemically reducible. In this case the problem of mental causation fades away. Under non-eliminative reductionism, the mental is a real (physical) thing in the world with physical causal properties. As such, the mental reliably causes behavior.

Conclusion

Criminal responsibility requires the truth of mental causation: a defendant is found criminally responsible for an act where she possesses certain culpable mental states that are causally related to criminal harm. In this chapter, I have attempted to review various theories regarding the relationship between mind and body to identify the theory that is the best candidate to support criminal responsibility assessments by articulating a convincing account of mental causation. Eliminativists can be seen as falling into two categories: Churchland-style eliminativism claims mental states do not exist such that they cannot cause behavior; the new eliminativists denies a type of mental causation they then claim is vital to criminal responsibility. Popular non-reductive theories disconnect the

mental from the physical/casual world such that the mental is ephiphenomenal (Searle-style non-reductionism) or can no longer have reliable causal effects (Davidsonian non-reductionism). Thus both types of eliminativism, and non-reductive physicalism, threaten criminal responsibility because they do not provide satisfactory accounts of mental causation.

If we assume the widely accepted positon of ontological physicalism, non-eliminative reductive physicalism about the mind offers the most plausible account of the full-fledged mental causation criminal responsibility requires. This is because it is clear on a non-eliminative reductive account of mind how folk mental state concepts carry causal properties. A non-eliminative reductivist may hold that folk concepts pick out a disjunction of local physical states, and picks them out reliably enough for us to consistently use such concepts to predict and understand human behavior. In any particular case, a mental state has causal properties as a particular token instantiation of this local disjunction.

The criminal law uses folk notions to seek mental states falling within categories or types of mental states within defendants. Criminal courts determine whether specific mental states can be assigned to a defendant and whether these states are causally related to the criminal harm. Such folk concepts seem to refer to conscious mental states, which in turn pick out particular token instances of a local disjunction of physical states. In general, it seems that folk explanations of human behavior are better positioned than scientific explanations to ground responsibility assessments. However, it may be that widespread damage or dysfunction to a particular brain area, connection, or process can be understood to impact an agent's ability to produce or review a particular mental state.

NEUROLAW AND RESPONSIBILITY FOR ACTION

Concepts, Crimes, and Courts

Edited by

BEBHINN DONNELLY-LAZAROV

University of Surrey

WITH DENNIS PATTERSON AND PETER RAYNOR

CAMBRIDGE
UNIVERSITY PRESS