Libertarian Free Will and the Argument from Reason

Angus Menuge
Concordia University Wisconsin
Angus.Menuge@cuw.edu

1. Introduction.

The argument from reason is really a family of arguments to show that reasoning is incompatible with naturalism. Here, naturalism is understood as the idea that foundationally, there are only physical objects, properties and relations, and anything else reduces to, supervenes on, or emerges from that. For our purposes, one of the most important claims of naturalism is that all causation is passive, automatic, event causation (an earthquake automatically causes a tidal wave; the tidal wave responds passively): there are no agent causes, where something does not happen automatically but only because the agent exerts his active power by choosing to do it. The most famous version of the argument from reason is *epistemological*: if naturalism were true, we could not be justified in believing it. Today, I want to focus on the *ontological* argument from reason, which asserts that there cannot be reasoning in a naturalistic world, because reasoning requires libertarian free will, and this in turn requires a unified, enduring self with active power.

The two most promising ways out of this argument are: (1) Compatibilism—even in a deterministic, naturalistic world, humans are capable of free acts of reason if their minds are responsive to rational causes; (2) Libertarian Naturalism—a self with libertarian free will emerges from the brain. I argue that neither of these moves works, and so, unless someone has a better idea, the ontological argument from reason stands.

2. Compatibilism and Human Rationality.

The basic idea of compatibilism is that a decision is free if it derives from rational causes. This assumes that reasoning is compatible with determinism. On Dennett's view, you are unfree if your actions result from a closed program, like the Sphex wasp that can be made to repeat the same actions indefinitely (move a cricket to the threshold of its burrow; go inside to check if it is safe) by moving the cricket away from the threshold when it is inside (it never just drags the cricket in, but moves it back to the threshold and goes inside to check if it is safe again).[1] What's wrong with the Sphex is that it is insensitive to the obvious fact that its routine is pointless, and can't break out of the loop. However, being controlled isn't the problem: what matter is *what* controls you: you are free so long as your will is governed by the right (rational) causes. Thus, a demonic neurologist might rob you of freedom by inducing irrational beliefs and desires, but if we were overwhelmed by the persuasive arguments of a well-informed truthful oracle, we would still be free. So long as reason drives the bus, we can be free even if, like Luther, we could do no other.[2]

A major problem for compatibilist theories of reasoning is that they don't tell us why some reasoning belongs to, or is owned by, a particular agent. The occurrence of phenomena responsive to rationality is not enough for reasoning: a notepad may be responsive to rational formulae, but it isn't reasoning; likewise a computer is responsive to a rational algorithm, but it is not reasoning for itself. This objection is standardly pressed through manipulation arguments, e.g., couldn't a kinder, gentler neurologist implant reasoning of his own in a subject? The subject is now responsive to reasoning, but his decisions are controlled by the neurologist's reasoning, not his own. This appears to show that a person's *using* reasons is not enough to show that he is reasoning for himself.

---

[1] Daniel Dennett, *Elbow Room: The Varieties of Free Will Worth Wanting* (Cambridge, MA: MIT Press, 1984), 10-13.

[2] Of course, Dennett's interpretation of Luther's claim is highly implausible. Not only could Luther have wimped out for prudential reasons, he was reporting the *result* of making up his mind, not the process of doing so.

In large part to address this sort of worry, J. M. Fischer and Mark Ravizza wrote their seminal work, *Responsibility and Control*.[3] Their key claim is that subjects are responsible for their decisions when they issue from a mechanism that is (1) moderately reasons-responsive and (2) the agent's own.

To be moderately reasons-responsive requires two things. First, you need a regular pattern of reasons-receptivity, which means that there is an intelligible pattern in the reasons someone would recognize in actual/possible cases: probably there is something a bit wrong with someone who would sell their Packers/Bears tickets for $1,000, but not for $2,000, $3,000 etc. Second, you need at least weak reasons-reactivity (that is, in at least some relevant possible worlds, one's decisions do reflect ones reasons): the requirement is made weak, because a weak-willed person who usually ignores his best reasons can still be held responsible. To address manipulation arguments, Fischer and Ravizza require that the mechanism yielding the decisions be the agent's own: this requires that the agent take responsibility for the mechanism by (1) seeing himself as the source of the decisions, (2) accepting he is a fair target for the "reactive attitudes" (praise and blame), (3) basing (1) and (2) appropriately on evidence. So, someone could take responsibility for decisions based on the oracle by (1) seeing himself as the source of the decisions because he endorsed the oracle's reasons, making those reasons his own, (2) accepting praise or blame for this decision, and (3) doing so on the basis of appropriate evidence (e.g., evidence that he wasn't coerced). In that case, a decision based on the oracle is the agent's own. But, it is claimed, one hasn't taken responsibility for unknown neural interventions as a source of one's own reasoning, and cannot be held responsible for any decisions made on that basis.

3. Critique of the Compatibilist Account of Human Rationality.

A. Reasons-responsiveness.

A basic problem is that "reasons-responsiveness" is ambiguous between a passive and an active notion. On the passive view, a computer is reasons-responsive because it responds to a rational program. However, all this shows is that the computer behaves *in accordance with reason*. Almost no-one thinks that the computer is reasoning for itself. So what compatibilists need is the active notion of reasons-responsiveness, according to which an agent actively selects, endorses, or takes responsibility for reasons, making those reasons his own. However, the problem is that only the passive notion of reasons-responsiveness is available to the naturalist. For on the naturalistic view, all an agent is reduces to, or depends on, a bundle of passive, automatic, event causal processes. There is no mental substance, agent cause or transcendent self over and above these processes that can select/endorse/take responsibility for some of them and not others. On the naturalistic view, we precisely are passive, organic computers, and so the best we can hope for is that our brains behave in accordance with reason, not that we can reason for ourselves or be responsible for our decisions.

Indeed, before we can talk of being responsible for our decisions, we need an account of why those decisions *belong* to us. But the trouble is, on a naturalistic view, there is no entity that can plausibly own any mental states, there is simply a plurality of parallel, impersonal processes in the brain. To see the problem, consider two examples.

Case 1. Suppose that Albert believes that A = B and also that B = C, and as a result of these two beliefs concludes that A = C. Albert's reasoning makes sense if there is some one mental substance (Albert) which owns and unifies the beliefs that A = B and B = C in one consciousness, and which endures over the time it takes to draw the conclusion that A = C. But if there are no mental substances, then even if one brain process contains the information that A = B and another contains the information that B = C, there is no entity that unites the information in one act of thought at a time, or which can persist

[3] John Martin Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (Cambridge: Cambridge University Press, 1998).

to draw a logical conclusion over time. Should the information that A = C arise in some later brain process, there is nothing which could be credited with having *reasoned* to that process. Without a unified, enduring subject, compatibilist naturalists have no way to distinguish Albert's act of reasoning from many other causal processes which would produce a rational conclusion without reasoning.

Case 2. Suppose that Phil (an athlete in an 8am logic class) believes that A = B and that B = C, but is so lethargic that he does not draw the obvious conclusion. Happily, a kind neurologist monitoring Phil's brain sees the information there and induces him to believe that A = C. Then Phil has reached a conclusion in accordance with reason, but he has not done so by reasoning. Fischer and Ravizza can respond that Phil did not take responsibility for the process that produced his belief, but this does not help the naturalist, because naturalism does not explain the existence of any entity that could take such responsibility. For if taking responsibility simply reduces to (or supervenes on, or emerges from) another brain process, then it must be possible for the neurologist to trigger that process as well, making it impossible to distinguish genuine cases of taking responsibility from brainwashing.

B. Unstoppable manipulation and the ratiomaniac.

This last point can be developed to show that, even with its many ingenious conditions designed to block cases of manipulation, Fischer and Ravizza cannot avoid all manipulation arguments. Suppose Jill, a logic professor and former neuroscientist, is exasperated with her weak student Jack. So, Jill implants a rationality enhancer in Jack's brain which not only makes Jack moderately reasons-responsive but also induces him to take responsibility for the decisions based on the rationality enhancer, by making him endorse the enhancer as part of himself, so that he sees himself as the source of the decisions it produces and an apt target for any praise or blame that ensues, all based on the apparent evidence that the enhancer is fully integrated with himself. In this case, Jack satisfies all of Fischer and Ravizza's conditions for responsibility. But arguably, although he thinks he is reasoning for himself, he is not. Should the implant contain a bug, Fischer and Ravizza would have to say incorrectly that Jack was responsible for the error of reasoning.[4] But not only is Jack not responsible (perhaps Jill is), *he* is not reasoning at all: he is simply the host of a rational parasite that has successfully coerced his endorsement of the parasite's "decisions" as his own, rather like individuals who have been converted by the cybermen on the science fiction show *Doctor Who*.

But perhaps the implants scenario seems artificial. No matter, the same problem can be raised without appeal to implants. It is perfectly conceivable that a human being suffers from compulsive moderate reasons-responsivity and also compulsively takes responsibility for his decisions. In his classic work, *Orthodoxy*, G. K. Chesterton unkindly compares the modern materialist to a maniac who has lost everything except his reason[5] and so continues thinking along the same trajectory even if it would serve his own best goals to re-consider. Call such a person a *ratiomaniac*. In the ratiomaniac's mind, one might say that reasoning *occurs* and carries on automatically, yet it seems wrong to say that the ratiomaniac is reasoning because the ratiomaniac has no power to stop thinking along this trajectory. A good example might be R. M. Hare's "paranoid man," who thinks obsessively about the possible dangers of every apparently kind action: "Mightn't the cupcakes be poisoned?"; "You are just being nice to lull me into a false sense of security," etc. Does that mean the paranoid ratiomaniac isn't moderately reasons-responsive? Not necessarily. It is perfectly possible that the paranoid ratiomaniac's pattern of decision-making follows a coherent, understandable pattern (regular reasons receptivity) and that at least sometimes he makes decisions on the basis of the reasons (at least weak reasons reactivity). Isn't it perfectly predictable that the paranoid man will think of possible hidden dangers and threats (which

[4]Some philosophers may be willing to bite the bullet, insisting that if determinism is true, one must be able to make someone responsible. Yet the intuition that Jack is not responsible for bugs in the implant seems stronger.

[5]See "The Maniac," chapter 2 of "G. K. Chesterton's, "Orthodoxy," in *The Collected Works of G. K. Chesterton*, vol. I (San Francisco, CA: Ignatius Press, 1986).

need not be logically absurd or excluded by known fact), and make precautionary decisions on that basis? And isn't it also possible that the ratiomaniac takes responsibility for his decisions, and for example, isn't surprised when people berate him for being "negative," "gloomy," "Puddleglum," etc, or when security professionals thank him for spotting potential dangers and security holes at airports? So on Fischer and Ravizza's account, the paranoid ratiomaniac is responsible for his decisions, yet it may be that he is in the grip of a highly sophisticated neurotic program, and is no more reasoning for himself, than is the obsessive compulsive individual who finds himself checking yet again if the door is locked.

Basically, the endemic problem with all compatibilist accounts of reasoning is that no matter how sophisticated we make the requirements, it is possible they are implemented in a Sphexish way, whether through conditioning, neural intervention or a natural psychological defect. It seems to me this reflects two obvious facts: that computer programs can automate virtually any rational procedure without reasoning, and that, for compatibilism, the brain just is an organic computer.

C. A non-Humean self.

John Searle argues that a Humean bundle of causal processes cannot account for human reasoning. He argues that we must postulate an irreducible non-Humean self, unified at a time and persisting over time.[6] Searle points out that in clear cases of human reasoning, an agent's reasons cannot be viewed as sufficient event causes of his decisions, for then there is no distinction between compulsive and non-compulsive decisions.[7] For example, we judge that if I am so seized by a desire, e.g. for double-chocolate cake, that I devour it like a machine, then although there was a reason for my action (my desire), my action is not the result of *reasoning*. Likewise if the ratiomaniac makes decisions that arise automatically from his reasons prior to deliberation, then deliberation is a pantomime that makes no difference to those decisions. Deliberation involves attending to evidence, goals and means, evaluating all of these, and drawing a practical or theoretical conclusion. But as Searle argues, this process has a point only if my beliefs and desires are not by themselves sufficient to yield the decision: there is a gap between my reasons and that decision (and other gaps as well[8]). If selves were just Humean bundles of beliefs and desires, then they would transition automatically to a decision. Since there is gap between these beliefs and desires and the decision, this gap must be bridged by something else, and the clear evidence of introspection is that this entity is a unified, enduring self, that owns these beliefs and desires, and evaluates and selects some of them in making that decision. The obvious problem of this admission, however is that a non-Humean self sounds like a mental substance, with active agent causal power, both of which are incompatible with a naturalistic ontology.

D. Libertarian Free Will.

Not only must this self be something over and above the bundle of its reasons, it must also have libertarian free will. If determinism is true, then the bundle of reasons prior to the decision must passively compel that decision. If it does not, as the evidence for a gap argues, then the self must have active power to select and endorse some of those reasons, thereby making the decision. In the case of logical reasoning, it is notorious that people do not try to draw all the logical conclusions of their beliefs, and aside from a lack of energy, this is partly explained by a willful dislike of where certain arguments are going. Famously, Thomas Nagel suggests that his own "fear of religion…is not a rare condition and…is responsible for much of the scientism and reductionism of our time,"[9] leading people to reject

---

[6]Searle, *Rationality in Action* (Cambridge, MA: MIT Press, 2001). He gives an extended argument for this conclusion on pages 79-96.
[7]Searle, *Rationality in Action*, 12-17.
[8]Searle also discusses the gap between the decision and the action (as when a student decides to get up for an 8am statistics class but lacks the willpower to follow through), and the gap between the initiation and continuation of an extended action (as when one tires of removing toilet paper from trees decorated by teenagers).
[9]Thomas Nagel, *The Last Word* (New York: Oxford University Press, 1997), 130-131.

disturbing conclusions that seem otherwise well-founded in their own reasoning. For example, suppose that Paul came to believe that *if* (A) [reasoning requires a transcendent self with libertarian free will], *then* (B) [naturalism is false], but after reading Searle's *Rationality in Action* was horrified to discover that he now believed that A. But Paul does not want to conclude that B (naturalism is false), so he converts the *modus ponens* into a *modus tollens*, claiming to have learned from the undeniable truth of naturalism that transcendent selves with libertarian free will are perfectly compatible with naturalism after all. The fact that we can willfully reject the conclusion our best reasons are pointing to is good evidence that those reasons do not compel us to accept a conclusion.

If Searle is right, there is no way to make sense of human reasoning without postulating an irreducible, non-Humean self with libertarian free will. Since these ontological commitments are *prima facie* outrageously non-naturalistic, one could be forgiven for concluding that Searle must finally have abandoned naturalism. But like our friend Paul, Searle has chosen to convert the *modus ponens* into a *modus tollens*, and, with the help of a sophisticated account of emergence[10], concludes that naturalism can accommodate the ontological demands of reasoning after all.

4. Searle's Libertarian Naturalism.

Searle says many things that make him sound like a dualist (and, I believe he really is), defending the reality, causal efficacy and ontological irreducibility of consciousness, intentionality, and more recently, libertarian free will. However, Searle denies that he is a dualist, on the grounds that "All of our mental states are caused by neurobiological processes in the brain, and they are themselves realized in the brain as its higher-level or system features."[11] Rather than offering a profound philosophical solution to the mind-body problem, Searle offers a combination of two responses. First, he claims that the philosophical problem can be dismissed, once we remove the confusions engendered by dualist vocabulary: we can see that the mind is causally, though not ontologically, reducible to the brain. Second, he claims that the hard problem is therefore not philosophical, but neurological: what is it about the brain that explains the emergence of mental phenomena?

Searle claims that the existence of such a self can be reconciled with naturalism by appeal to the idea of emergence. Although he admits some limitations to the analogy, Searle thinks Sperry's account of the emergence of a wheel's solidity has many parallels with the emergence of the self from the brain.

> The wheel is entirely made of molecules. The behavior of the molecules causes the higher-level, or system feature of solidity… [T]he solidity affects the behavior of the individual molecules. The trajectory of each molecule is affected by the behavior of the entire solid wheel. But…there is nothing there but molecules…we are not saying that the solidity is something *in addition* to the molecules; rather, it is just the *condition* that the molecules are in. But the feature of solidity is nonetheless a real feature, and it has real causal effects.[12]

By analogy, Searle claims, in a conscious human brain, "the behavior of the neurons is causally constitutive of consciousness" and "the neuronal structures move [the] body…because of the conscious state they are in. Consciousness is a feature of the brain in the way that solidity is a feature of the wheel."[13] One major disanalogy between the two cases which Searle allows is that consciousness, unlike solidity, "is not ontologically reducible to physical microstructures," due to its first-person ontology.[14] However, he claims that consciousness is still causally reducible to the brain, in the sense that it "has no causal powers beyond the powers of the neuronal (and other neurobiological) structures."[15]

---

[10]Emergentists do not have to be naturalists, of course. William Hasker endorses an emergent dualism. I will focus my remarks on emergent naturalism, which is ably defended by Timothy O' Connor, David Hodgson, and Robert Kane as well as John Searle.
[11]John R. Searle, *Freedom and Neurobiology* (New York: Columbia University Press, 2007), 40.
[12]John R. Searle, *Freedom and Neurobiology*, 48-49.
[13]John R. Searle, *Freedom and Neurobiology*, 49.
[14]John R. Searle, *Freedom and Neurobiology*, 50.
[15]John R. Searle, *Freedom and Neurobiology*, 50.

In this setting, Searle proposes that we can give a neurobiological translation of the problem of free will: "what would the behavior of the neurons…have to be like if the conscious experience of free will were to be neurobiologically real?"[16] He considers two hypotheses.

On **Hypothesis 1**, there are gaps at the psychological level, but there are none at the neurobiological level, so the experience of freely acting in the gaps is an illusion. But as Searle notes, Hypothesis 1 implies epiphenomenalism, the implausible view that our mental lives make no difference to our behavior, and is insufficient to ground rational decision-making, since all decisions are caused and realized by brain states which arose automatically from previous brain states.

On **Hypothesis 2**, however, "the absence of causally sufficient conditions at the psychological level is matched by an absence of causally sufficient conditions at the neurobiological level."[17] Hypothesis 2, he claims, could account for rational decision-making. Suppose that t1 denotes some time at which deliberation is occurring and t2 denotes a time at which a decision is made. Then, Searle's account requires that three conditions are met:

> First, the state of the brain at time t1 is not causally sufficient to determine the state of the brain at t2. Second, the movement from the state at t1 to the state at t2 can only be explained by features of the whole system, specifically by the operation of the conscious self. And third, all of the features of the conscious self at any given instant are entirely determined by the state of the microelements, the neurons, etc. at that instant.[18]

Thus on Searle's account there is: (1) diachronic indeterminism (later states of consciousness/neurons are not determined by prior states); (2) apparent top-down causation (system-level features have a causal impact on micro-level features); and (3) synchronic determinism (at any given time, each conscious state is determined by its realizing neuronal state). Searle is, however, clearly uncomfortable with the idea of top-down causation, affirming on several occasions that there are no gaps in the brain, and hoping to deflate the problem by claiming that talk of levels is a misleading metaphor, because "Consciousness is located in certain portions of the brain and functions causally, relative to those locations."[19] While acknowledging that mere randomness does not account for rational decision-making, Searle speculates that if the brain is a quantum system, this might help to explain how higher-level system features like consciousness can act back on the neural structures that cause them.[20] So long as the higher-level system features are purely physical, causal closure is not violated.

5. Critique of Searle's Libertarian Naturalism.

A. The Location Problem.

In his recent work, J. P. Moreland reminds us of Frank Jackson's distinction between two ways of doing metaphysics.[21] One way, is the *shopping list approach to metaphysics*. On the shopping list approach, one collects items for one's ontology because they make life easier (they help to make one's theories work), or simply because one likes them. One problem with this approach is that no serious attempt is made to show that these items are compatible with one's ontology. In *serious metaphysics*, one must *locate* the items in one's preferred ontology, by showing why they are a better fit with that ontology than they are with rival ontologies.

In brief, the problem with Searle's solution to the problem of human reasoning is that he combines a shopping list approach to the ontology of rational deliberation with mysterious hand-waving

---

[16] John R. Searle, *Freedom and Neurobiology*, 58.
[17] John R. Searle, *Freedom and Neurobiology*, 62.
[18] John R. Searle, *Freedom and Neurobiology*, 64-65.
[19] John R. Searle, *Freedom and Neurobiology*, 63.
[20] John R. Searle, *Freedom and Neurobiology*, 74-75.
[21] Frank Jackson, *From Metaphysics to Ethics* (Oxford: Clarendon Press, 1998). Moreland appeals to Jackson's account of the location problem in his *Consciousness and the Existence of God* (New York: Routledge, 2008) and his *The Recalcitrant Imago Dei* (London: SCM Press, 2009).

about emergence. He admits that human reasoning makes no sense unless there is a non-Humean self, capable of unifying its thoughts at a time and persisting over time, but his explanation of why any such marvelous entity exists is simply that the brain happens to cause it. However, this makes the connection between the brain and the self merely contingent, and a contingent relation between the mental and the physical is compatible with dualism. So Searle's account fails to show that physicalism provides a better explanation than dualism of the mental-physical correlation. The contingency problem is standardly brought out by appealing to thought experiments[22]: surely there are conceivable worlds which satisfy all of the same physical descriptions, but in which there is no consciousness, intentionality or rationality, or in which experiences and thoughts are redistributed in bizarre ways (e.g. inverted spectrums, Jeff experiencing torture while his body is massaged, Sue feeling relaxed when she is being physically tortured, etc.). Further, as Searle employs it, "emergence" is simply a loaded label for the correlations that need explaining, not an explanation of those correlations.

Searle is aware of the contingency argument, but argues that the relationship between the mental and the physical is more than contingent, asserting that, given the laws of nature and the physical facts, it is not possible that the mental facts would be different.[23] However, as Stewart Goetz and Charles Taliaferro point out, this argument is either irrelevant or it commits an obvious modal fallacy.[24] Even if it is true that *given* the laws of nature and the physical facts, the mental facts are necessary, this does not show that the physical facts necessitate the mental facts, because the laws of nature are not themselves necessary. The laws of nature may include high-level descriptions of mental-physical correlations that always hold in our world, but there may be other worlds bound by quite different laws in which those correlations do not obtain. As a result, Searle fails to show that the non-Humean self required for human reasoning is best explained by a naturalist ontology and so he does not solve the location problem.

B. The Exclusion Problem.

Searle's brand of naturalism qualifies as a form of non-reductive physicalism, in the sense that he claims that the emergent self and its properties cannot be ontologically reduced to the physical. However, Jaegwon Kim has argued persuasively that non-reductive physicalism faces the problem of explanatory exclusion.[25] Non-reductive physicalists claim that although mental states cannot be identified with physical states (by either type or token identity), mental states are nonetheless entirely dependent on physical states in a way that physical states are not dependent on mental states (e.g. supervenience, emergence). The non-reductive physicalist hopes to affirm both mental causation and the ontological priority of the physical over the mental. However, Kim shows that the core principles of physicalism are not compatible with mental causation, and hence that non-reductive physicalism is an unavailable option.

To see this, consider any case of mental causation. Suppose mental state M causes a further mental state M*. By hypothesis, M is completely determined by some physical base state P, and M* is completely determined by some physical base state P*. Given the assumed priority of the physical over the mental, M* cannot exist without its base P* (or some alternative base which we may assume is not present), so M must cause M* by causing P*. However, physicalism is also committed to the causal closure of the physical which implies that every event has a purely physical cause. So, given the dependence of M on P it is natural to say that P causes P*, and hence that P cause M*. But, assuming

---

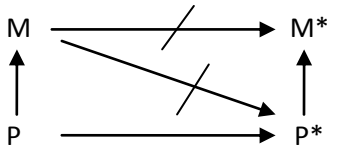
[22]See J. P. Moreland, *The Recalcitrant Imago Dei*, 30-31.

[23]John Searle, *Mind: A Brief Introduction* (New York: Oxford University Press, 2004), 128-129.

[24]Stewart Goetz and Charles Taliaferro, *Naturalism* (Grand Rapids, MI: Eerdmans, 2008) 77-78.

[25]Jaegwon Kim has made this case against non-reductive physicalism in many places, including his *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation* (Cambridge, MA: MIT Press, 1998), his *Philosophy of Mind*, second edition (Cambridge, MA: Westview Press, 2006) and his essay "Causation and Mental Causation," in eds. Brian P. McLaughlin and Jonathan Cohen, *Contemporary Debates in Philosophy of Mind* (Malden, MA: Blackwell, 2007).

we do not allow systematic overdetermination, if P causes M*, and P has ontological priority over M, then M cannot also be the cause of M*: M is excluded. Generalizing, granted physicalism, there can be no distinctively mental causation.



KEY:
1) Vertical arrows signify supervenience.
2) Horizontal arrows signify causation.
3) Crossed arrows are excluded causal pathways.
M is excluded by P, which causes M* by causing P*.

So, Kim concludes, the physicalist really has only three options: (a) *epiphenomenalism*—the mental exists but makes no causal contribution; (b) *eliminativism*—the mental does not exist; or (c) *ontological reductionism* (Kim's preferred solution)—mental states just are physical states, so mental states can cause things, but nonetheless all causation is physical causation.

We know, however, that Searle rejects all three of these alternatives, and it might seem that his account has two resources for avoiding the exclusion problem: (1) indeterminism; and (2) subtle differences between emergence and supervenience. However, on inspection, neither of these resources can defeat the exclusion problem.

(1) Indeterminism.

Searle claims that the state of the brain at t1 does not determine the state of the brain at t2, and that, within this causal gap, a non-Humean self is able to reason to a decision. He suggests that the brain may be a quantum system, so that the transition from one brain state to the next may have a probability of less than 1. However, indeterminism does not by itself solve the exclusion problem. For on a physicalist view, all causation must reduce to the passive event causation recognized by paradigmatic physical theories. If indeterminism is allowed, the only difference is that causes fix the chances of their effects, rather than necessitating those effects. This means that there is no room for the active power of a non-Humean self to *alter* the chances of an effect, and so no room for this self to make a distinctive causal contribution.

(2) Emergence.

Might not the self emerge from neuronal structures in the way that the solidity of the wheel emerges from the wheel's molecules? Then we could say that just as solidity is nothing but a condition the wheel's molecules are in, and yet acts back on those molecules (e.g., by making a difference to their location when the wheel is rolling), so the self's consciousness is a condition that neurons are in, and yet acts back on those neurons by closing the gap between an agent's reasons and his decisions. However, this move still does not evade the exclusion problem.

For one thing, as Moreland points out, the analogy between solidity and consciousness does not work. Solidity is a structural property that emerges *of necessity* when the molecules are arranged in a certain way, and Searle admits that "the solidity of the wheel is ontologically reducible to the behavior of the molecules."[26] There is therefore nothing that solidity can add to what a full microphysical account of the molecules already explains: there is no gap to fill and so solidity is excluded by the microphysical account from having any *additional* or *distinctive* explanatory role. One can continue to say that solidity does things, but this is (at best) useful but redundant shorthand for the microphysical account. Now either consciousness is a structural property of the brain or it is not. If it is[27], then, like solidity, it is

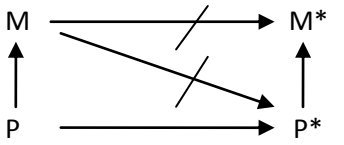
[26]Searle, *Freedom and Neurobiology*, 64.
[27]It quite obviously is not, because consciousness is not an aggregate of separable parts.

excluded from having a distinctive causal role. If it is not, then the problem is that since consciousness is unlike solidity and any of the other standard physical examples of emergence (such as liquidity, digestion and photosynthesis), the claim of emergence remains unmotivated. In none of these examples do we find the emergence of states essentially characterized by a first-person ontology. So these analogies give no reason to think that physical emergence is a more plausible account than dualism. Therefore, it seems that the only way in which Searle can avoid the exclusion problem is to embrace the substance of dualism, but call it emergence.

What is more, Kim explicitly shows how to adapt his exclusion argument to undercut emergentist solutions to the mind-body problem.[28] Suppose that emergent mental state M causes emergent mental state M*. For example, a conscious self affirms some reasons and not others (M) and this produces a rational decision (M*). According to emergentism, M* is synchronically necessitated by its physical base state P*. It is also reasonable to assume that if M* is multiply realized, no alternative base state is actually present. So, in the closest worlds in which P* is absent, M* would not happen. But if so, the only way M could cause M* is by causing P*. Thus, if emergentism is true, mental-to-mental causation requires downward (mental-to-physical) causation. However, as an emergent state, M must also have a physical base P. But then we have a problem:

> If causation is understood as nomological (law-based) sufficiency, P, as M's emergence base, is nomologically sufficient for it, and M, as P*'s cause is nomologically sufficient for P*. It follows that P is nomologically sufficient for P* and hence qualifies as its cause. The same conclusion follows if causation is understood in terms of counterfactuals… .[29]

But if P qualifies as the cause of P*, and hence M*, unless we allow systematic overdetermination, M is excluded from any causal role. But this is *déjà vu*.



KEY:
1) Vertical arrows signify *emergent determination*.
2) Horizontal arrows signify causation.
3) Crossed arrows are excluded causal pathways.
M is excluded by P, which causes M* by causing P*.

C. The Epiphenomenalism Problem.

That Searle's emergent physicalism is in an inconsistent bind can be shown by contrasting his reasons for introducing a transcendent self with his physicalist qualms about causation. Searle asserts that we need to postulate a self to bridge the gap between causally insufficient reasons for an action and a decision, which implies that the self makes a causal contribution. However, as Moreland points out, although Searle explicitly rejects epiphenomenalism, his naturalistic commitments imply it.[30] For on Searle's account, emergent properties can have no causal powers beyond those of the underlying physical processes (causal reductionism). Therefore if mental properties do emerge, they cannot alter what these physical processes were already going to do (downward causation is impossible). But then, mental properties are epiphenomenal, and a transcendent self cannot bridge any causal gaps.

D. The Substantial Selves Problem.[31]

Although Searle grants that a non-Humean, transcendent self is required to explain human reasoning, he rejects substantial agent causes as not only "mistaken philosophy" but "bad English,"[32]

---

[28]Jaegwon Kim, "Emergence: Core Ideas and Issues," *Synthese* (2006) 151: 547-559.
[29]Jaegwon Kim, "Emergence: Core Ideas and Issues," *Synthese* (2006) 151: 558.
[30]J. P. Moreland, *The Recalcitrant Imago Dei*, 64.
[31]In this section I do advance several positive considerations in favor of mental substances, but I do not claim to offer any direct refutation of rival forms of dualism, such as Hasker's emergent dualism. Whether such rival theories can do as well or better in explaining reasoning is a question I leave for another day.

because in a genuine causal explanation we must say what it is about an object that accounts for the effect. But not only is this argument pedantic, it also overlooks the obvious dualist reply that what it is about substantial selves that enables them to cause things is their possession of active power, the ability to initiate and redirect causal chains. Further, Searle seems quickly to forget his own qualms about agent causation when he describes how selves bridge the gap. His whole reason for wheeling in the self is that an agent's reasons are typically causally insufficient to produce a decision. But if the self actually contributes something that was not going to happen anyway, it seems this can only be by its exercise of active power, which requires a substantial self. As we have seen, this cannot be avoided by appeal to event causal indeterminism, since then the self cannot alter the chances fixed by its brain states. And without a substantial self, Searle cannot solve the problem of compulsive rationality, because the self cannot alter what those reasons were going to do anyway unless it is capable of downward causation, but this is excluded by naturalism.

Further, a substantial self appears to be required to account for those characteristics of a self which Searle himself agrees are required to explain human reasoning.

First, that self must exhibit *unity at a time*, uniting all of the self's reasons. But such unity is not credible on the basis of a naturalistic ontology. The self is supposed to emerge from underlying brain processes, but these processes are massively parallel, distributed across many different regions of the brain, and are in constant flux. If the brain processes realizing an agent's reasons are in different areas of the brain and proceed independently, there is no identifiable physical entity that can be said to unify all of those reasons (this is called "the binding problem" for materialism). Nor is it plausible that the unifying entity emerges from those processes. For, as Moreland points out, physical systems can be fully understood as aggregates of *separable parts*: the parts are not ontologically dependent on the wholes and can exist without those wholes.[33] But an agent's reasons are *inseparable parts*: the parts cannot exist without the whole. Thus Jack's belief that Harvard is overpriced is inseparable from Jack himself. Although Jill can have a belief with the same content, Jill cannot have Jack's belief, because that belief is intrinsically tied to Jack and cannot exist without him. So an agent's reasons are not the kind of thing that can be understood in physical terms, as separable parts of an aggregate. But reasons do make sense as inseparable parts if they are modes (property-instances) of a substance, in the way, for example, that *this* piece of gold's malleability (that property instance) is inseparable from the piece of gold.

Second, the self must *persist over time*, so that it is the same entity which has reasons that draws conclusions from them. Yet on the physicalist account, the neural processes exhibit a passive succession of events in constant flux. There is no identifiable brain state that persists over time and which could ground a persistent, emergent self. This point is confirmed by neuroscientist Mario Beauregard, who has extensively studied the neural correlates of consciousness:

> No single brain area is active when we are conscious and idle when we are not. Nor does a specific level of activity in neurons signify that we are conscious. Nor is there a chemistry in neurons that always indicates consciousness.[34]

Even supposing the binding problem could be solved, all one could hope for is that from each successive brain state, a transient self, or "I-stage" would emerge.[35] But if transient self I1 emerges from brain state B1, and B1 causes some quite different brain state B2 from which transient self I2 emerges, there is no reason to suppose that I1 and I2 are the same self. To merely assert that the event causal flux of the brain processes happens to give rise to a self that remains identical over time is both implausible

---

[32]John Searle, *Rationality in Action*, 82.

[33]Moreland, *The Recalcitrant Imago Dei*, 109.

[34]Mario Beauregard and Denyse O'Leary, *The Spiritual Brain: A Neuroscientist's case for the Existence of the Soul* (New York: HarperCollins, 2007), 109.

[35]This argument is indebted to one of J. P. Moreland's arguments against Timothy O' Connor's emergentism. See Moreland's *The Recalcitrant Imago Dei*, 141.

and non-explanatory. If the identity of the self cannot be grounded at the neurological level, then no reason has been given to prefer a physicalist account to dualism, and so identity has not been located in a naturalistic ontology. On the other hand, if the self is a substance in its own right, and has essential qualities that remain despite its causal interactions with the world, it is not hard to see how it could persist over time. So again, a substantial self gives a better explanation of how reasoning is possible.

Third, Searle is clear that we need selves to explain acting for a reason because rationality presupposes an entity with *libertarian free will* that can select and act on some reasons rather than others. But the problem with this is that the underlying naturalistic ontology does not permit the existence of libertarian free will or the ability to act on reasons. Given the exclusion argument above, it is impossible for the self to make a difference to what the neurological realization of the agent's reasons was already going to do. But even if the exclusion argument could be avoided, it would still be incompatible with physicalism to say that the self acts on reasons. This is because reasoning is inherently teleological, but physicalism does not permit goal-directed causal processes.

Fourth, reasoning requires that intentional content plays a causal role, but naturalism cannot countenance irreducible intentionality. Suppose that Tom believes that A = B and that B = C. It is the content of those beliefs (plus Tom's goal of deriving the right conclusion) that explain Tom's inferring that A = C. For suppose that we asserted instead that only physically respectable properties can causally explain anything. Then we would have to say something like this: the fact that the belief that A = B is realized in brain state B1 and that the belief that B = C is realized in brain state B2 explains the occurrence of a brain state B3 that realizes the belief that A = C. This might be a causal explanation, but the way in which beliefs are realized in the brain gives no reason for the conclusion, so it is not a rational explanation. Given the exclusion argument, however, a naturalist must offer just such a causal explanation, and so cannot capture why Tom's inference is rational. In other words, for naturalism precisely what makes something a reason (its intentionality) is epiphenomenal, so while an individual might happen to transition to a brain state realizing a conclusion that is in accordance with reason, he cannot reason *to* that conclusion. By contrast, if Tom is a mental substance intrinsically characterized by subjectivity, intentionality, teleology and active power, he can inspect his own internal, intentional states—his reasons—and make an inference which satisfies his goal of deriving the conclusion best supported by those reasons.

6. Conclusion.

The ontological argument from reason purports to show that reasoning cannot occur in a world exclusively populated by a naturalistic ontology because it requires a unified, enduring self with libertarian free will. To avoid this argument, the naturalist can make one of two main moves. Like Dennett, Fischer and Ravizza, he can attempt to show that some version of naturalistic compatibilism suffices to account for reasoning, because even in a world of event causation, some creatures may be responsive to reason. But, amongst many other problems, such an account fails to distinguish compulsive rationality that is merely occurring in someone's brain from reasoning that an agent does. The vital distinction between acting in accordance with reason and acting from reason appears to require the postulation of libertarian free will. So, like Searle, the naturalist may attempt to show that libertarian free will can be naturalized. But on inspection, even sophisticated notions of emergence do not overcome the problems of location, exclusion and epiphenomenalism. If the emergent self bridges the explanatory gap by adding something new the account is not naturalistic. But if the account is faithful to naturalistic ontology, the self cannot bridge the explanatory gap. So given the strength of Searle's arguments for an irreducible self, some form of dualism seems unavoidable to account for reasoning, and, though I have not attempted to exclude several rival dualist accounts, I have argued that substance dualism is up to the job.