Journal of Philosophical Theological Research
Print: 1735-9791
Online: 2538-2500

AI, Persons, Free Will, and a Policy Recommendation

John Lemos

McCabe Professor of Philosophy, Department of Philosophy, Coe College, USA.

jlemos@coe.edu

Abstract

This essay examines whether artificial intelligence (AI) could ever qualify as persons entitled to the same rights and responsibilities as human persons. The discussion engages two classic contributions to the philosophy of mind: William Lycan's argument that a suitably programmed machine could possess mentality and free will, and John Searle's Chinese Room argument, which casts doubt on the claim that syntactic symbol manipulation alone can generate genuine understanding or intentionality. The author concurs with Lycan that a properly programmed AI could, in principle, possess free will, but advances a distinct account of this possibility grounded in an event-causal libertarian theory of free will. The author argues that he and Robert Kane have both advanced theories of libertarian free will that could allow a properly programmed AI to possess free will. Nevertheless, the author contends that Searle's critique furnishes substantial grounds for skepticism regarding the personhood of any future AI, since genuine understanding and first-person intentionality may not be realizable in computational systems. Given this residual uncertainty, and given the profound moral and legal implications that would follow from granting personhood to AI, the essay proposes a precautionary policy recommendation: we should refrain from creating AI systems that, for all we know, may or may not be persons. The argument thus integrates metaphysical analysis of free will and personhood with applied ethical and policy considerations, offering a framework for responsible development of increasingly sophisticated artificial agents.

Original Research

Keywords:

artificial intelligence, personhood, free will, libertarianism, moral status, AI policy.

Received: 2026/03/12; **Received in revised form:** 2026/05/15; **Accepted:** 2026/06/08; **Published online:** 2026/06/12

Lemos, J. (2026). AI, Persons, Free Will, and A Policy Recommendation. *Journal of Philosophical Theological Research*, (Special Issue on AI, Consciousness, and Human Agency), 28(3), 53-76. https://doi.org/10.22091/jptr.2026.17031.3639

Copyright © The Author

Introduction

We typically think that most adult human beings have capacities for conscious understanding – conscious knowledge – of themselves and their surroundings. We also think the typical adult human being has powers of reason and that they can act rationally in ways that make them responsible for what they do. And we typically think that to be responsible for what we do – or, at least, to be responsible in such a way that we could deserve praise or blame for our actions – we must also possess free will.

I believe it is possible that one day there may be AI that possesses such powers of conscious understanding, reason, and free will. I also believe that if we knew such AI existed, then we should regard them as persons and grant them the same rights we grant to human persons. In exploring the prospects for there one day being conscious, rational AI that possesses free will, I consider two old, classic essays by William Lycan and John Searle (Lycan, 1987/2013, pp. 358–363; Searle, 1980/2013, pp. 351–358). In the essays I discuss, they both express an openness to the possibility of there one day being AI that are persons, though Searle, as we will see, is – and I think rightly – more skeptical about this.

I start by discussing key points in Lycan's essay. I argue that he is correct in thinking that were there ever robotic AI who acted for all the world in the way that human beings do, then we should regard them and treat them as persons. He maintains that AI could even possess free will, and in doing so he explains how they could meet soft-determinist standards for free-willed action. While I agree that AI could one day possess free will, I give reasons to reject his soft-determinist account, and I go on to argue that there is no reason in principle why AI could not possess libertarian free will. In later parts of this article, I discuss Searle's essay, noting the grounds of his more skeptical view of the prospects for AI personhood.

Unlike Lycan, Searle is more concerned that the power for conscious understanding may be tied to having the kinds of human brains that we have. Searle thinks it may be that no man-made AI that lacks a human brain may be capable of any conscious understanding.

I argue that his deeper skepticism about the prospects of AI personhood is warranted, but I nonetheless maintain that if we get to a point where AI are indistinguishable from human beings in their behavior patterns, then we would be morally obligated to nonetheless treat them as persons with all of the rights and responsibilities of human persons. I conclude by arguing that such considerations morally justify a ban on making AI beings that blur the lines between persons and nonpersons.

William Lycan's More Optimistic Vision of the Prospects for AI Personhood

In his essay, William Lycan describes a fictional robot, Harry, who looks and acts like many human beings. In the following passage, he describes Harry and raises some interesting questions about him:

> Harry (let us say) is humanoid in form – he is a miracle of miniaturization and has lifelike plastic skin – and he can converse intelligently on all sorts of subjects, play golf and the viola, write passable poetry, control his occasional nervousness pretty well, make love, prove mathematical theorems (of course), show envy when outdone, throw gin bottles at annoying children, etc., etc. We may suppose he fools people into thinking he is human. Now the question is, is Harry really a *person*? Does he have thoughts, feelings, and so on? Is he actually conscious, or is he just a mindless walking hardware store whose movements are astoundingly like those of a person? (Lycan, 2013, p. 359)

Clearly, Harry is not a human being; he is not a member of the biological species *Homo sapiens*. But might he not, nonetheless, be a person, that is, a being with the kind of conscious understanding of himself and the world and powers of will and reason that make him deserve full moral consideration and, thus, make him deserving of rights to life and liberty which we ought to respect?

It is Lycan's view that were we ever to encounter a being like Harry, then we should regard him as a person and treat him with the same moral respect that we should show toward human beings. While Harry is not a human being and not made of the organic, biological materials we are made of, we should nonetheless assume he is a person and deserves our moral respect. In support of this view, Lycan talks about how we come to believe that other human beings besides ourselves are persons deserving of our moral respect. He notes that to each one of us it is self-evident that we have conscious understanding of ourselves and the world and powers of will and reason. For example, it is self-evident to me that these things are true of myself, and I presume it is self-evident to you, my dear readers, that these things are self-evidently known by you of your own selves. But, as Lycan claims, what is not self-evident to me, nor to any of you reading this, is that there are any other beings that have conscious understanding of themselves and the world and powers of will and reason. Rather, each one of us comes to know of the existence of other persons through rational inference.

It is Lycan's view that, to know that other human beings have conscious understanding of themselves and the world and the powers of will and reason

– to know that other human beings are persons – we must infer this based on the behavior of other human beings. The idea here is that when I have certain beliefs and desires, I act in certain ways, and when I see other human beings acting in the ways that I act when I have such beliefs and desires, I have reason to believe that they have similar beliefs and desires. For instance, when I am hungry and there is no food in the house, then I go out to secure some food to eat, typically by going to the grocery store. Because I act this way in such circumstances, then when I see that my wife hasn't eaten in the past twelve hours, and I see her look in the empty refrigerator and cupboards, and I see her walking down the street toward the market, I presume that she feels hungry, she knows there's no food in the house, and she is going to get some food. I presume she has similar kinds of conscious feelings, knowledge, and powers of reason that lead her to will her movement toward the market for food.

According to Lycan, since this is how we establish that there are human beings other than our own selves that have conscious understanding and powers of will and reason, and since Harry the robot also behaves in the ways that we ourselves would act given our beliefs, desires, feelings, and reason in similar circumstances, we then have good reason to believe that he, too, has powers of conscious understanding, will, and reason. That is, we have good reason to believe that he, like ourselves and other human beings, is a person. Lycan states:

> We know that a human being has such-and-such mental states when it behaves, to speak very generally, in the ways we take to be appropriate to organisms that are in those states. . . . We know for practical purposes that a creature has a mind when it fulfills all the right criteria. And by hypothesis, Harry fulfills all our behavioral criteria with a vengeance. . . . It follows that we are at least *prima facie* justified in believing him to be conscious.

We have not *proved* that he is conscious, of course – any more than you have proved that I am conscious. An organism's merely behaving in a certain way is no logical guarantee of sentience; from my point of view it is at least imaginable, a bare logical possibility, that my wife, my daughter, and my chairman are not conscious, even though I have excellent, overwhelming behavioral reason to think they are. But, for that matter, our "standard behavioral tests" for mental states yield practical or moral certainty only so long as the situation is not palpably extraordinary or bizarre. (Lycan, 2013, p. 360).

He goes on to note that many will think, and rightfully so, that the situation with Harry is "palpably extraordinary or bizarre," as he acts just like a human being, but he is not a human being. Thus, some will balk at the suggestion that he is a person with powers of conscious understanding, reason, and will.

Lycan notes that what makes Harry's case different is the fact that he is not born of human parents (rather, he was made in a lab) and that he is not made of the same things as human beings (rather, he is not made of carbon compounds but of silicon). But Lycan goes on to argue that these facts don't give sufficient reason to think that he lacks personhood, that he lacks the powers of conscious understanding, reason, and will.

I see no obvious way in which either a creature's origin or its sub-neuroanatomical chemical composition should matter to its psychological processes or any aspect of its mentality. (Lycan, 2013, p. 360)

With this in mind, he famously tries supporting this intuition with the example of Henrietta, who is born as a human being and who, as an adult, has her body parts – arms, legs, heart, even parts of her brain – gradually replaced with inorganic parts over a long period of time. He has us suppose that, eventually, over a long stretch of time, all her body parts, including all her brain parts, are replaced with inorganic (nonbiological) parts, but in the process "her intelligence, personality, perceptual acuity, poetic abilities, etc., remain just as they were before" (Lycan, 2013, p. 360). Were this to happen, it seems we ought to regard Henrietta as enduring through all this change and having all the same rights and responsibilities she had when she was composed of biological parts.

Interestingly, before concluding his essay, Lycan goes on to contemplate a possible objection to his view that nonbiological machines could conceivably one day rightly be regarded as persons. In doing so, he notes that some might say that if such machines are rightly considered to be persons, then they should have free will. He says it could be argued that, since any such machines would operate on programming given them by their makers, they would lack free will. The worry is that if all of the behavior of a machine, like Harry, is just the determined result of programming placed in him by his programmer, then he would lack the kind of free will that some might think is essential to personhood.

Now, Lycan could have simply rejected the notion that having free will is an essential ingredient of personhood. But he doesn't. Instead, he argues that there are ways in which machines, like Harry, could rightly be thought to have free will. In support of this, he argues that, for all we know, soft determinism about free will might be true, and if so, then the fact that Harry always acts and chooses in accordance with his programming does not establish that he lacks free will. Lycan says soft determinism is the view that "to have freedom of choice in acting is (roughly) for one's action to proceed out of one's own desires, deliberation, will, and intention, rather than being compelled or coerced by external forces regardless of one's desires or will" (Lycan, 2013, p. 361). Thus, on this conception of free will, even if all of Harry's desires, deliberations, willings, and intentions are the determined result of the way he is programmed

by his designers and his environment, he would still act with free will as long as his actions are caused by those desires, deliberations, willings, and intentions. After all, while his thoughts, desires, and deliberations may all be determined by the complex interplay of his programming and environmental influences, those thoughts, desires, and deliberations are still *his*, and if he acts in accordance with his desires, thoughts, and deliberations, then, on the soft determinist view, he is acting freely. Harry would only act unfreely when external forces inhibit his capacity to act in accord with his thoughts, desires, and deliberations. Since he embraces such a soft-determinist perspective on the nature of free will, Lycan sees the issue of free will as posing no threat to the idea that machines like Harry could be persons. Thus, he thinks that maybe one day in the future such nonbiological machines may come into existence, and if they do, we should treat them as persons, granting them rights to life and liberty.

There are two points I want to make about Lycan's essay. First, while I agree with Lycan that robots like Harry and Henrietta could possess free will, I don't think he has given an adequate account of how this is possible. Second, I agree that if robots, like Harry and Henrietta, came into existence, then we should treat them as persons, granting them all the same rights to life and liberty that we grant to human persons. However, unlike Lycan, I believe there are sufficient reasons to doubt the personhood of Harry and Henrietta, and I believe these reasons for doubt have moral implications for whether we should allow for the construction of such robots. In the next section of this essay, I will discuss issues of free will, arguing that soft-determinism is false and that even on a correct alternative theory of free will, the libertarian theory, it is conceivable that one day there may be robots that meet libertarian standards of free will. After that, I will consider ideas and arguments from John Searle which raise significant doubts about whether any robots will ever meet the conditions of personhood, and I will discuss some of the moral implications of these lingering doubts.

Free Will and Robots

In explaining how robots can have free will, Lycan appeals to a soft-determinist conception of free will. However, there are good reasons to reject such approaches to free will. On such a view of free will, when a person's actions are causally determined by their own beliefs and desires and deliberations, and are not the products of external force or coercion, then they act with free will. For instance, if I want to go into a swimming pool because I think it would be pleasant, and I believe it would not be wrong of me to go into the pool, and these beliefs and desires causally determine me to go into the pool, then I go into the pool of my own free will. But if I have no such beliefs

and desires, and if a strong wind blows me into the pool while I'd rather not go in, then I do not go into the pool of my own free will. If the soft-determinist theory of free will is true, then even if everything we ever do is a causally determined consequence of prior states and events, we can still have free will.[1] And, as noted, Lycan believes that since soft-determinism about free will is true, robots like Harry and Henrietta can have free will. That is, he believes that even if robots receive their programming from the engineers and programmers who design them, and even if all their behaviors are the determined result of the complex interplay between their programming and the situations they face in their environment, they could still rightly be said to possess free will.

I reject such views of free will. To have free will, we must have the kind of control over our lives and actions which grounds deserved praise for good acts and good moral character and deserved blame for bad acts and bad moral character. But meeting soft-determinist standards of free will does not suffice for this, and this is because we cannot have desert-grounding free will if everything that we ever do is causally determined. Rather, to have the kind of free will that is the ground of desert, some of our actions must be causally undetermined free-willed acts. Notice I say that *some* of our actions must be causally undetermined free-willed acts if we are to have free will. It is my view that many of our free-willed acts may be causally determined by our character or motivational states, but to be free-willed, such determined acts must issue from a character, a self, we have formed by prior causally undetermined free-willed acts. Only if this is the case can we possess desert-grounding responsibility for such causally determined actions. As did the great libertarian theorist of free will, Robert Kane, let us call these causally undetermined free-willed acts which shape our character, our self, "self-forming acts" (SFAs).

To see why free will requires the ability to form oneself through causally undetermined free-willed actions – to see why the capacity for SFAs is essential to free will – suppose that Harry has been programmed with certain

1. In contemporary philosophy, soft-determinists are typically understood as believing that (a) free will is compatible with determinism and (b) determinism is true. Thomas Hobbes ([1651] 1958), David Hume ([1748] 1974), and John Stuart Mill (1874) were soft-determinists. But in contemporary philosophy, not as many philosophers are committed to the truth of determinism, while many nonetheless still believe that even if determinism were true, or even if all human actions were determined, we could still have free will. Such philosophers tend to refer to themselves as "compatibilists." For some recent and contemporary defenses of compatibilism, see A. J. Ayer (1954), Daniel Dennett (1984a, 1984b, 2003), Owen Flanagan (2002, ch. 4), Harry Frankfurt (1969, 1971), G. E. Moore (1912), Kai Nielsen (1971), J. J. C. Smart (1963), Gary Watson (1975), and Susan Wolf (1987, 1990). John Martin Fischer defends what he calls "semi-compatibilism," the view that while free will is not compatible with determinism, moral responsibility is. See Fischer (1994, 1999, 2006) and Fischer and Ravizza (1998).

sorts of values. Perhaps he is programmed with a belief in the Kantian formula of the end, which says we ought always to treat persons as ends in themselves and never as mere means. Thus, in any given situation involving the treatment of persons, Harry always treats them as ends and never as mere means. Thus, he never lies, nor cheats, nor steals, nor rapes, nor murders, because he is programmed to act in accord with the principle of ends. He follows these Kantian norms even when he knows that he could get away with and profit from violating these norms, and, again, he follows these norms because this is how his programming operates. Suppose Harry always does the right thing as a result of his being given this Kantian programming, which determines that he will always make the ethical choice. Would he deserve our praise for always doing the right thing? Many people would, I think, say, "No, he does not deserve our praise, because all of his right actions are just the necessary consequence of the programming that has been bestowed upon him." And I agree with this sort of reaction. Furthermore, if Harry had been programmed with a bad moral framework which necessitated his acting badly in many situations, then I do not think he would deserve our blame either.

Now, if I am right that having free will grounds deserved praise and blame for our actions, then soft-determinism is misguided. For if soft-determinism is true, then Harry should be said to act with free will whenever he acts in accordance with his beliefs, desires, and deliberations. But this would entail that he acts with free will in acting with the Kantian morals he has been programmed to have and which he necessarily follows. Yet, it seems intuitively obvious that he doesn't deserve praise for all the good actions he does in this manner, nor would he deserve blame if he had been programmed to act badly. Thus, if it's true that free will grounds deserved praise and blame, then soft-determinism seems to be false.

To have desert-grounding free will, agents – and Harry included – must be responsible for shaping their own moral characters, their own selves. And to do this, they cannot act in determined ways from programs or characters imposed upon them solely by external factors, whether by programmers, enculturation, genetics, or the like. Rather, there must be moments in their lives in which they make causally undetermined free choices which have an impact on shaping their future selves for better or for worse. The sort of view that I am defending here is a libertarian view of free will of the sort that Robert Kane and I have defended over the years in many of our writings (Kane, 1996, 2005; Lemos, 2018, 2021, 2023a, 2023b). My own views are largely Kaneian in nature; however, we offer slightly different views about the nature of self-forming acts.[1] I would note that Kaneian

1. For expressions and defenses of my own modified view of self-forming acts (SFAs), see my (2018, Ch.5, esp., pp.104-112, 2021, 2023a, Ch. 6, 2023b, 2025a, 2025b). For critical

views of free will raise certain sorts of questions that must be answered to make the theory plausible. In what follows, I will say more about this and then move on to talk about what must happen in a robot if it is ever to act with desert-grounding free will. In discussing these issues further, I will just focus on Kane's view, not my own variant of Kane's view.

Kane maintains that, to be responsible for our characters – to deserve praise or blame for our characters and for the actions causally determined by our character – we must be responsible for causally undetermined actions which shape our character. For instance, if I have a bad character which frequently leads me in a causally determined way to do bad things, then, on Kane's view, to be responsible for those causally determined bad acts and to deserve blame for those bad acts, I must have contributed in significant ways to forming that bad character through prior causally undetermined free-willed acts. Many philosophers are skeptical about whether a person can have desert-grounding control over causally undetermined actions. In support of this skepticism, it has been argued that if an action is causally undetermined at the moment of choosing, then two persons, P and P*, in different possible worlds facing the same choice between two options, A and B, could choose differently even if everything about them, including all their actions and psychological states, had been exactly the same until the moment right before their choice. But if their life histories and their psychological states had been exactly the same throughout their entire lives right up to the moment before they choose, and they choose differently, then nothing about them explains the difference in their choices. As such, their different choices seem to be unexplainable random happenings, mere matters of luck, suggesting that they would lack desert-grounding control over such causally undetermined choices.[1]

Kane famously maintains that we can be responsible for undetermined actions, including undetermined choices. In support of this point, he appeals to Austin-style examples, so named after the famous philosopher J. L. Austin, who originally presented such examples. Kane has us consider an assassin who is trying to shoot and kill a prime minister. He has us imagine that, as the assassin takes aim and is pulling the trigger, it is causally undetermined whether the assassin will have a nervous twitch making his shot go off course, leading

discussions of my modified view of SFAs, see Caruso (2025), Haji (2022, 2023) and Kane (2024, Ch.7, pp.175-181), Pereboom (2025), and Sekatskaya (2025).

1. This possible worlds version of the luck objection has, perhaps, been most forcefully defended by Alfred Mele (1998, 1999a, 1999b). Another famous version of the luck objection is Peter van Inwagen's "replay" or "rollback" argument (2002). For other expressions of luck objections, also see Allen (2005), Caruso (2021), Clarke (2002), Haji (1999, 2005), Levy (2011), Pereboom (2001, 2007a, 2007b, 2014), Strawson (2000), Waller (1988).

him to miss the prime minister. Kane says suppose there is no nervous twitch and the assassin succeeds in killing the prime minister. In such a case, the assassin's killing of the prime minister would be causally undetermined, but he would still be responsible for killing him. And, according to Kane, he would be responsible for killing the prime minister because he would have succeeded in what he was trying to do. When we succeed at what we are trying to do, we will be responsible for what we do even if what we do is causally undetermined.

According to Kane, this helps explain how we can be responsible for causally undetermined free choices. To see this, let's first consider the kinds of conditions that give rise to the making of causally undetermined self-forming choices. Kane maintains that these kinds of choices occur when we are torn between two or more different possible courses of action, A and B, and we have some significant motivation to do either one. These are the kinds of situations that lead us to deliberate prior to choosing which course of action to take. In the most recent expression of his view,[1] Kane maintains that during deliberation, as the agent finds a certain option, say A, more attractive, the agent will likely start making an effort of will – start trying – to choose A. However, since he is strongly attracted to option B as well, it is causally undetermined whether the agent's effort to choose A will succeed. If that first effort to choose A succeeds, leading him to choose A, then his choice will be causally undetermined, but he will be responsible for it, because his subsequent choice of A will be a result of his effort. However, if his initial effort to choose A fails, then he may either stop deliberating and not choose, or he may deliberate further, possibly making another effort to choose A or making an effort to choose B. Suppose he makes an effort to choose B. It will then be undetermined whether this effort succeeds, because he still has some significant attraction to choose A. But if this later effort to choose B succeeds, then, again, he will be responsible for his choice of B, as it is something he tried to choose.

On Kane's view, when making these undetermined choices, whatever the agent chooses will be a result of his effort to make that particular choice. With our example in mind, if he chooses A, that will be the result of an effort to choose A; if he chooses B, that will be a result of his effort to choose B. And since, as the Austin-style examples demonstrate, we are responsible for

1. In very recent years, Kane has changed his understanding of SFAs. He used to believe that in SFAs the agent simultaneously makes distinct efforts of will to choose each option considered in deliberation. Due to criticisms of this point, he now rejects the idea that agents engage in such simultaneous conflicting efforts of will. For expressions of the relevant criticisms of Kane's old view, see Balaguer (2010, 179n23), Caruso (2021, 49), Clarke (2003, 88), Ekstrom (2003), and Lemos (2018, 73-81, 2023a, 119-127). For expressions of Kane's more recent view, which is designed to avoid these problems as well as the problem of luck, see Kane (2021, 2024).

undetermined acts when we succeed at what we are trying to do, we are then responsible for causally undetermined free choices in cases like those described above.

Furthermore, it is Kane's view that in our lives we often face these choices when we are torn between doing what is morally right and doing what is wrong, or when we are torn between satisfying some present desire when we would be better off waiting and pursuing some other course of action. And he believes that, depending on how we choose in these sorts of cases, it contributes to the formation of our character, our selves, contributing in significant ways to whether we are morally decent persons or morally bad persons or, for that matter, weak-willed or strong-willed. The idea is that if I commit too many self-forming acts in which I make bad choices, then I am likely contributing to my having a morally bad character. Thus, when in later years I have a bad character and this determines me to act badly, then I am responsible for those later bad actions because I have contributed in significant ways to making myself morally bad.

I am very sympathetic with this picture of the nature of free will and responsibility. While I have developed a slightly different view about the nature of self-forming acts, replacing Kane's efforts of will directed at particular choice options with the indeterministic assigning of evaluative weight to different choice options, I find the overall framework of his theory and the grounds for his rejection of soft-determinist and compatibilist views of free will to be correct. With these points in mind, let's consider what must be the case if a robot, an AI, is to have free will in the Kaneian sense. First, the robot must be constructed and programmed in such a manner that it has conscious understanding of itself and the world. It must be able to formulate goals and be able to think and reason and be consciously aware of its goals and its reasoning. Second, it must be able to make efforts of will to achieve its goals, since, on Kane's view, making such efforts of will is crucial to making causally undetermined free-willed choices. Third, there must be situations it faces in which the goals it has come into conflict, creating the need for deliberation and choice among these conflicting goals. And fourth, there must, of course, be causal indeterminacy at work in some of the practical reasoning – some of the deliberations and decision-making – of any robot that has free will.

Regarding the last point about causal indeterminacy in deliberation and decision-making, it is important to understand that, with respect to human beings, when an undetermined free-willed choice between two options, A and B, is made, it is not sufficient for the agent to make efforts for the different choice options and then merely have one effort succeed merely due to a randomly occurring neural blip. Rather, Kane has always been clear that the causal indeterminacy of neural

functioning that allows for undetermined free choices opens up a window in which an agent can make one of his efforts of will overcome resistance, allowing the agent to make the choice to which that effort was directed. So, for instance, if I choose A and not B, the indeterminacy in neural processing merely leaves it open as to whether my effort to choose A will succeed. The choice of A must not be seen merely as the result of some random neural blip that causes that choice. Rather, the indeterminacy in neural functioning merely allows that my effort to choose might or might not succeed, and when the effort does succeed, the choice is a result of that effort, and I am responsible for the choice. As such, the indeterminacy merely opens up a window in which I can make, in an undetermined way, one of my efforts succeed at making a choice, thereby establishing my responsibility for the choice.[1] I stress this point here to make clear that in any AI that has free will, it would not suffice for its undetermined choices *merely* to be the product of random decision -output generators. Were this the case, the choice would not be free-willed. Rather, for an AI to make a free-willed undetermined choice of A instead of B, the AI must make an effort of will to choose A, and it must be able to make that effort succeed in the face of resistance brought on by its recognition that option B is also a valued goal that it has.

---

1. As some evidence of this aspect of Kane's view, I would note how Kane has replied to the idea that on his view our undetermined free-willed decisions would exert no more free will than having our decisions decided by random decision output generators. Dennett (2003, pp. 132-133) has made such an objection, but he notes that in correspondence Kane replies:

   > The indeterminacy-producing mechanism [leading a person to make an undetermined free-willed choice] must be responsive to the dynamics within the agent's own will and not override them or it would be making the decisions and not the agent. [Quoted from Kane in Dennett (2003, 133)]

   Elsewhere when discussing SFAs, Kane writes:

   > [O]ne needs to say more than that certain reasons (for doing A) win out in the process of deliberation at the moment of choice, or certain other reasons (for doing B) win out, and it is undetermined which ones win out. One must also be able to say that the agent makes one set of reasons win out over the other at the moment of choice, so that the agent can be fully responsible for causing it to be the case that one choice rather than the other is made, despite the indeterminism. And this requires more resources.
   >
   > To do it, I argue, the activity of the agent prior to choice must not merely be imagined to be a single process (deliberation) with multiple possible outcomes (deciding to A or deciding to B), one of which occurs by chance. One must also consider that there is an internal complexity to the preceding process of deliberating that gives the agent more of an active role in settling which reasons win out over the others. This can be done if, instead of merely imagining a single process (deliberation) with multiple possible goals, one also imagines that within that larger process, there are multiple processes, each with a specific goal the agent is trying to realize. These processes, of course, are my efforts of will or volitional streams…
   >
   > In such a scenario, it would not merely be the case that one set of reasons "wins out" over the other, but rather that the agent makes one set of reasons prevail over the other by making an effort to do so against the competing effort to make a contrary choice. (Kane, 2014, p. 208).

Kane's theory of free will is a naturalistic theory of free will. He makes no appeal to Kantian noumenal selves or Cartesian nonphysical souls to make sense of free will. Rather, on his view and mine, human beings are complex material beings. Our thinking, understanding, reasoning, and willing arise from the operations of our brains. Thus, if the complex material beings that we are can engage in causally undetermined free-willed decision-making, then I see no reason *in principle* why there may not someday be machines, AI, that can do the same. This, of course, is not to say that there are any currently existing AI that have the kind of free will that makes them responsible for what they do. I think we are far from that.

In the next section, I will address some epistemological concerns about whether we could ever really know that we are in the presence of AI persons deserving of rights to life and liberty. Before doing so, I want to address one other nuance about AI free will, responsibility, and deserved praise and blame. Suppose that a robot, like Harry, has just been made. And suppose in making him he was programmed to have: (i) certain goals and certain moral beliefs; (ii) a significant amount of understanding about his own self and the world around him; (iii) the ways and means to increase his understanding about himself and the world in which he lives; and (iv) the ways and means to rationally deliberate in the face of any situations creating conflicts in his goals. What I want to note here is that Kane and I would say that when first brought into existence, Harry would have very little responsibility for anything that he does. Recall that, on the view I have put forward, we only engage in causally undetermined free-willed choices when we encounter conflicts in our will, and much of what we do is not done in the face of conflict. Rather, much of what we do is the determined result of the kind of character we have and our motivational states. Thus, when first brought to life and released into the world, Harry will mostly engage in causally determined acts programmed into him by his programmers. He would not be responsible for such causally determined acts. Rather, his programmers would be responsible for those acts. But in time, as he navigates his way through the world, if he faces conflicts in achieving his goals and must reason through and make undetermined choices through efforts of will, he may then slowly build up increasing levels of responsibility for those undetermined choices and for the character shaped by these choices. Over the years, as he makes more and more causally undetermined free-willed choices, he will increasingly contribute to his own self-formation, and this will make him increasingly responsible for his character and the determined choices he makes as a result of his character in the later stages of his existence.[1]

1. For an alternative recent defense of the possibility of AI free will, see Christian List (2025). Like Kane and myself, List thinks to have the power of free will agents must be able to make

John Searle's More Skeptical Thoughts on the Prospects for AI Personhood

In another famous old essay, John Searle presents an argument which raises significant doubt about whether there could ever be AI that have conscious understanding of themselves, their environment, and what they are doing. This is his classic Chinese Room argument. This argument proceeds by analogy. Searle has us imagine that he does not understand the Chinese language at all. He is put into a room with no windows, and there is one wall of the room where there is a slot through which written texts may be submitted. There is another slot where written texts may be sent out of the room. He has us also imagine that Chinese people are outside the room and that they are encouraged to submit Chinese written comments and questions into the room through the input slot. The Chinese people do this. Searle, inside the room, is supposed to provide responses to these written Chinese texts. In the room is a massive book with instructions telling the reader what Chinese texts to send out in response to any Chinese texts coming in. Searle uses these instructions to make replies to any Chinese texts which are sent into the room, and he sends out written Chinese responses through the output slot. The Chinese people outside receive the replies that Searle sends out. The replies that Searle sends out are perfectly sensible replies to any comments or questions submitted. For instance, if a message sent in Chinese into the Chinese room means "Hello, how are you?", Searle, using his response instruction book, sends out Chinese words which mean something like, "Fine, thank you. How are you?" But when Searle does this, he has no idea what his written responses say. In the room, Searle sends out all kinds of reasonable Chinese responses to Chinese comments and queries, but Searle, not understanding any Chinese, has no idea what the comments and queries coming in mean, and he has no idea what the comments and queries going out mean. The people outside the room will likely, and somewhat reasonably, conclude that there is someone inside the room who understands Chinese and understands their submitted comments and queries and understands the replies he sends out of the room. But they are mistaken, because Searle understands no Chinese at all.

Searle says that this is how computers operate. They receive inputs of data and give responses, but in doing so they have no understanding of the incoming

---

decisions while having alternative possibilities. But the way in which he makes sense of such alternative possibilities has been subjected, and I think rightly, to significant criticism. See Kane (2024, pp. 232-235) and Pereboom (2023).

For a more negative assessment of the prospects for AI free will, see Lynne Rudder Baker (1981), who argues that AI cannot act at all. A being that cannot act cannot have free will. For a critical assessment of Baker's view, see Berkich (2018). For a different account of the nature of agency and the prospects of AI agency, see Helen Steward (2012, esp. p. 91).

data, and they have no understanding of their responses to this incoming data. It is his view that computers are programmed to respond to data they encounter in the ways that their programmers want them to respond, but in doing this, the computers have no understanding of the incoming data nor of their responses to the data. Indeed, a highly skilled programmer might be able to program a computer to perfectly simulate an intelligent human being who can engage in long, extended conversations with a human interlocutor in such a way that, if the human being didn't know any better, he would think that he was talking with another human being. However, if a computer is programmed to do this, it would still not understand any incoming data or the meaning of any of its responses to this incoming data. For these reasons, Searle believes that no computer understands what it is or what it does, even though it may perfectly simulate human responses to its environment, including responses to questions and comments from human interlocutors (Searle, 2013, pp. 352–353).

To be clear, Searle's point is that computers process information – take in data and respond to data – by instantiating a program. And, given the right sort of program, they can *simulate* human consciousness and understanding, just as the set-up of the Chinese Room can simulate human consciousness and understanding of the Chinese language. But *merely* instantiating a program that simulates human consciousness and understanding does *not* suffice for actually being conscious and understanding (Searle, 2013, p. 355). According to Searle, computers, understood in this sense, can *never* possess consciousness and understanding. As such, computers, understood in this sense, could never be persons and could never possess free will nor be morally responsible for what they do, and we could have no obligations to any computers understood in this sense.

However, Searle is open to the possibility that an artifact – a man-made machine – can think. In answer to the question of whether a man-made machine can think, he writes:

> Assuming it is possible to produce artificially a machine with a nervous system, neurons with axons and dendrites, and all the rest of it, sufficiently like ours, again, the answer to the question seems to be, obviously, yes. If you can exactly duplicate the causes, you could duplicate the effects. And indeed it might be possible to produce consciousness, intentionality, and all the rest of it using some other sorts of chemical principles than those human beings use. It is, as I said, an empirical question. (Searle, 2013, p. 355)

But he continues, writing:

> "Ok, but could a digital computer think?"

> If by "digital computer" we mean anything at all that has a level of description where it can correctly be described as the instantiation of a computer program, then again the answer is, of course, yes, since we are the instantiation of any number of computer programs, and we can think.
>
> "But could something think, understand, and so on *solely* by virtue of being a computer with the right sort of program? Could instantiating a program, the right program of course, by itself be a sufficient condition of understanding?"
>
> This I think is the right question to ask, though it is usually confused with one or more of the earlier questions, and the answer to it is no.
>
> "Why not?"
>
> Because formal symbol manipulations by themselves don't have any intentionality; they are quite meaningless; they aren't even symbol manipulations, since the symbols don't symbolize anything. (Searle, 2013, p. 355)

Interestingly, Searle says, "Yes, a digital computer can think." But in reply to the question, "Could instantiating a program, the right program of course, *by itself* be a sufficient condition of understanding?" [my emphasis added], he says "No." His point is that what a thing is made of, or at least how it is physically organized, is crucial to whether it can be conscious, whether it can have thought and understanding. Just because there are currently computers that can be programmed to fool us into thinking that they are persons with understanding is not sufficient evidence that they actually do think and understand. According to Searle, for something to think and understand, to have genuine consciousness, it must be made of the right sort of material and organized in the right sort of way.

The passages from Searle which I quote above suggest that he is open to there being something materially different from us which has consciousness. He says, "[I]t might be possible to produce consciousness, intentionality, and all the rest of it using some other sorts of chemical principles than those human beings use. It is, as I said, an empirical question." But in the end, he suspects that the capacity for conscious awareness and understanding is a biological phenomenon. In a later passage in the same essay, he writes:

> Could a machine think? My own view is that only a machine could think, and indeed only very special kinds of machines, namely brains and machines that had the same causal powers as brains. . . . Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be causally dependent on the specific biochemistry of

> its origins as lactation, photosynthesis, or any other biological phenomena. (Searle, 2013, p. 357)

In this classic essay, Searle is, in essence, maintaining that: (1) mere excellent simulation of human intelligence by a digital computer does not suffice for actually possessing intelligence; (2) a nonbiological, man-made digital computer made of the right sorts of materials and organized in the right sort of way could possibly be intelligent; and (3) he harbors serious doubts that such a being will ever be made, because he suspects there is something about the biological material of which we are made that is essential to producing intelligence. In response to Lycan's story about Harry the Robot, Searle would harbor serious doubts about Harry's possession of any intelligence – conscious understanding – of what it is and its surroundings. Searle would believe it is most likely that Harry merely simulates understanding and personhood without possessing understanding and personhood, and this is because Harry is not made of the right sort of material that would give us good reason to think he has conscious understanding and personhood. Lycan, by contrast, would be much more open to the idea that Harry has conscious understanding of himself and the world and free will and that Harry is a person. In talking about Harry, as well as Henrietta, Lycan writes:

> We do not discriminate against a person who has a wooden leg, or a mechanical kidney, or a nuclear heart regulator; no more should we deny any human or civil right to Harry or Henrietta on grounds of their origin or physical makeup, which they cannot help. (Lycan, 2013, p. 360)

I would note, however, that even Lycan would harbor some level of skepticism about Harry's personhood. As I noted earlier, he says regarding Harry the robot, "We have not *proved* that he is conscious, of course – any more than you have proved that I am conscious. An organism's merely behaving in a certain way is no logical guarantee of sentience; from my point of view it is at least imaginable, a bare logical possibility, that my wife, my daughter, and my chairman are not conscious, even though I have excellent, overwhelming behavioral reason to think they are" (Lycan, 2013, p. 360). But despite this reservation about Harry, Lycan does not see this as good reason to think of him as a nonperson nor to deny him the rights that we grant to human persons.

Concluding Remarks: On the Ethics of Building AI like Harry and Henrietta

In this essay, I've discussed two old and famous essays that are relevant to the

question of whether there may one day be man-made, nonbiological AI that have the kind of conscious understanding, reasoning abilities, and free will that would make them persons. Both the essay by Lycan and the one by Searle express openness to this possibility, while Searle seems to be more skeptical about it than Lycan.

Along the way, I've said a good bit about the nature of free will and what it would take for an AI to have free will. I argued that it's conceivable that an AI could have the kind of robust libertarian free will that's essential to ground just deserts and responsibility. Some people may think having such free will is an essential feature of personhood and full-fledged membership in the moral community.

In closing, I would note that I believe that if we ever do make robotic AI, like Harry and Henrietta, who act for all the world like human beings and who would easily be mistaken for human beings, we should regard them and treat them as persons and grant them all the rights to life, liberty, property, and so forth that we grant human persons. However, as both Searle and Lycan point out, just because a being meets all our behavioral tests for being a person, this is no guarantee it is a person. This fact, coupled with the fact that man-made AI, like Harry and Henrietta, would lack the material, biological brains of human beings, justifies significant doubt about whether they would have the kind of conscious understanding of themselves and the world that is essential to having desert-grounding free will and personhood.[1]

Again, if we end up making AI that are like Harry and Henrietta, then we should grant them all the same human and civil rights as human beings. And why? Not so much because we'd know that they deserve such rights, but because we should err on the side of safety. It would be horrible to deny them rights to life and liberty if they have mental lives that include the capacities for conscious understanding, reasoning, hoping, dreaming, fearing, and free will that are the hallmarks of personhood. At the same time, however, and as noted, there would be significant reason to doubt their personhood despite their certainly seeming to have it. Furthermore, regarding them as persons with all the rights and responsibilities that come along with this, we would, as a society, have to grant significant amounts of time, money, and effort in seeing to it that they are cared for and protected in the ways we care for and protect human

1. Just to be clear, regarding the connection between conscious understanding and free will, I would note that if a being lacks any conscious understanding, it will not have the kind of libertarian free will that Kane and I endorse. For instance, on Kane's view to act with libertarian free will one must engage in efforts of will aimed at making particular choice options, and doing this requires conscious understanding of the options and understanding the aims of one's efforts.

beings. As noted, it would be horrible to not recognize them as persons with rights and treat them as such, if they *are* persons with rights. But it would also be a great waste of time, money, and effort to treat them as persons and to respect robot rights if, in reality, they are unconscious automatons. Thus, with all these points in mind, I do not think that human beings should make AI that may lead us to wonder if they are persons with rights. We should not make AI, like Harry and Henrietta. We don't need them, as the world has plenty of persons as is. But if we do make AI, like Harry and Henrietta, who, for all we know, are persons, then, to err on the side of safety, we should treat them as persons, even though this may be a costly waste of our time, efforts, and resources.

I am, in essence, suggesting that once we have AI, like Harry and Henrietta, which (who?) for all the world – act like human beings in all the ways that would suggest they are persons deserving of moral rights and respect, we will then still have legitimate uncertainties as to whether they really merit the granting of such rights and respect. The problem faced here is well discussed in an important jointly authored report, *Taking AI Welfare Seriously* (Long et al., 2024), which was supported by Eleos AI Research and NYU's Center for Mind, Ethics, and Policy. In Section 1.2, "Risks of Mishandling AI Welfare," of this report, the authors do a wonderful job of spelling out this problem, noting that it would be terrible to not respect the rights of AI if they really are persons, but it would also be a terrible waste of resources, time, and effort to grant them rights if they ultimately are more like mindless automatons. As I argue in this essay, I think we should not make such highly sophisticated AI so as to avoid getting caught in between this rock and this hard place. Of course, as a society, we will have to address issues regarding at what point an AI is sufficiently person-like to ban its construction. I have not addressed this issue here, but it is one that must be confronted.[1]

▣ Conflict of Interest

The author declares no competing interests.

1. For a really nice discussion of issues dealing with this, see Lawrence & Morley (2021).

References

Allen, R. F. (2005). Free will and indeterminism: Robert Kane's libertarianism. *Journal of Philosophical Research, 30*, 341–355. https://doi.org/10.5840/jpr20053042

Ayer, A. J. (1954). Freedom and necessity. In *Philosophical essays* (pp. 3–20). St. Martin's Press. https://doi.org/10.1007/978-1-349-00132-3_12

Baker, L. R. (1981). Why computers can't act. *American Philosophical Quarterly, 18*, 157–163.

Balaguer, M. (2010). *Free will as an open scientific problem*. MIT Press.

Berkich, D. (2018). Machine intentions. *APA Newsletter on Philosophy and Computers, 18*, 3–10.

Caruso, G. (2021). *Rejecting retributivism: Free will, punishment, and criminal justice*. Cambridge University Press.

Caruso, G. (2025). Event-causal libertarianism, indeterministic weightings, and the pragmatic approach. *Philosophia, 53*, 905–914. https://doi.org/10.1007/s11406-025-00813-0

Clarke, R. (2002). Libertarian views: Critical survey of noncausal and event-causal accounts of free agency. In R. Kane (Ed.), *The Oxford handbook of free will* (pp. 356–385). Oxford University Press.

Clarke, R. (2003). *Libertarian accounts of free will*. Oxford University Press.

Dennett, D. C. (1984a). *Elbow room*. MIT Press.

Dennett, D. C. (1984b). I could not have done otherwise – so what? *Journal of Philosophy, 81*, 553–567.

Dennett, D. C. (2003). *Freedom evolves*. Viking Press.

Ekstrom, L. (2003). Free will, chance, and mystery. *Philosophical Studies, 113*, 153–180. https://doi.org/10.1023/a:1023940209581

Flanagan, O. (2002). *The problem of the soul: Two visions of the mind and how to reconcile them*. Basic Books.

Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy, 66*, 829–839. https://doi.org/10.2307/2023833

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*, 5–20. https://doi.org/10.2307/2024717

Haji, I. (1999). Indeterminism and Frankfurt style examples. *Philosophical Explorations, 2*, 42–58.

Haji, I. (2005). Libertarianism, luck, and action explanation. *Journal of Philosophical Research, 30*, 321–340. https://doi.org/10.5840/jpr20053041

Haji, I. (2022). Libertarianism and luck. *Journal of Philosophical Theological Research, 24*, 115–134. https://doi.org/10.22091/JPTR.2022.8297.2723

Haji, I. (2023). Further reflections on Lemos's indeterministic weightings view of libertarian free will. *Journal of Philosophical Theological Research, 25*, 119–130. https://doi.org/10.22091/JPTR.2023.9722.2929

Hobbes, T. (1958). *Leviathan*. Bobbs-Merrill. (Original work published 1651)

Hume, D. (1974). An enquiry concerning human understanding. In *Locke, Berkeley, and Hume, The empiricists*. Anchor Books. (Original work published 1748)

Kane, R. (1996). *The significance of free will*. Oxford University Press.

Kane, R. (2002). Some neglected pathways in the free will labyrinth. In R. Kane (Ed.), *The Oxford handbook of free will* (pp. 406–437). Oxford University Press.

Kane, R. (2007a). Libertarianism. In J. M. Fischer, R. Kane, D. Pereboom, & M. Vargas (Eds.), *Four views on free will* (pp. 5–43). Blackwell.

Kane, R. (2007b). Response to Fischer, Pereboom, and Vargas. In J. M. Fischer, R. Kane, D. Pereboom, & M. Vargas (Eds.), *Four views on free will* (pp. 166–183). Blackwell.

Kane, R. (2011). Rethinking free will: New perspectives on an ancient problem. In R. Kane (Ed.), *The Oxford handbook of free will* (pp. 381–404). Oxford University Press.

Kane, R. (2014). New arguments in debates on libertarian free will: Responses to contributors. In D. Palmer (Ed.), *Libertarian free will: Contemporary debates*. Oxford University Press.

Kane, R. (2021). Making sense of a free will that is incompatible with determinism: A fourth way forward. *Journal of Philosophical Theological Research, 23*, 5–28. https://doi.org/10.22091/JPTR.2021.7247.2595

Kane, R. (2024). *The complex tapestry of free will: A philosophical odyssey*. Oxford University Press.

Lawrence, D., & Morley, S. (2021). Regulating the Tyrell Corporation: The emergence of novel beings. *Cambridge Quarterly of Healthcare Ethics, 30*, 421–434. https://doi.org/10.1017/S0963180120000973

Lemos, J. (2018). *A pragmatic approach to libertarian free will*. Routledge.

Lemos, J. (2023a). *Free will's value: Criminal justice, pride, and love*. Routledge.

Lemos, J. (2023b). Haji and the indeterministic weightings model of libertarian free will. *Journal of Philosophical Theological Research, 25*, 101–118. https://doi.org/10.22091/JPTR.2023.9648.2920

Lemos, J. (2025a). Precis of *Free will's value: Criminal justice, pride, and love*. *Philosophia, 53*, 905–914. https://doi.org/10.1007/s11406-025-00870-5

Lemos, J. (2025b). Replies to Caruso, Elzein, Pereboom, Sekatskaya, and Smilansky. *Philosophia, 53*, 973–993. https://doi.org/10.1007/s11406-025-00904-y

Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility*. Oxford University Press.

List, C. (2025). Can AI systems have free will? *Synthese, 206*, 115. https://doi.org/10.1007/s11229-025-05209-x

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). *Taking AI welfare seriously* (arXiv:2411.00986) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2411.00986

Lycan, W. (2013). Robots and minds. In J. Feinberg & R. Shafer-Landau (Eds.), *Reason and responsibility: Readings in some basic problems of philosophy* (pp. 358–363). Wadsworth. (Original work published 1987)

Mele, A. (1998). Review of Robert Kane's *The significance of free will*. *Journal of Philosophy, 95*, 581–584. https://doi.org/10.5840/jphil1998951122

Mele, A. (1999a). Kane, luck, and the significance of free will. *Philosophical Explorations, 2*, 96–104. https://doi.org/10.1080/13869799908520970

Mele, A. (1999b). Ultimate responsibility and dumb luck. *Social Philosophy and Policy, 16*, 274–293. https://doi.org/10.1017/S0265052500002478

Mill, J. S. (1874). *A system of logic*. Harper and Row.

Moore, G. E. (1912). Free will. In *Ethics*. Oxford University Press.

Nielsen, K. (1971). The compatibility of freedom and determinism. In *Reason and practice*. Harper and Row.

Pereboom, D. (2001). *Living without free will*. Cambridge University Press.

Pereboom, D. (2007a). Hard incompatibilism. In J. M. Fischer, R. Kane, D. Pereboom, & M. Vargas (Eds.), *Four views on free will* (pp. 85–125). Blackwell.

Pereboom, D. (2007b). Response to Kane, Fischer, and Vargas. In J. M. Fischer, R. Kane, D. Pereboom, & M. Vargas (Eds.), *Four views on free will* (pp. 191–203). Blackwell.

Pereboom, D. (2014). *Free will, agency, and meaning in life*. Oxford University Press.

Pereboom, D. (2023). Review of *Why free will is real*? *Criminal Law and Philosophy, 17*, 229–234. https://doi.org/10.1007/s11572-022-09634-8

Pereboom, D. (2025). Event-causal libertarianism without settling. *Philosophia, 53*, 961–972. https://doi.org/10.1007/s11406-025-00914-w

Searle, J. (2013). Minds, brains, and programs. In J. Feinberg & R. Shafer-Landau (Eds.), *Reason and responsibility: Readings in some basic problems of philosophy* (pp. 351–358). Wadsworth. (Original work published 1980)

Sekatskaya, M. (2025). Manipulation, indeterminacy, and the luck problem: On Lemos' *Free will's value*. *Philosophia, 53*, 953–960. https://doi.org/10.1007/s11406-025-00878-x

Smart, J. J. C. (1963). Free will, praise, and blame. *Mind, 70*, 291–306.

Steward, H. (2012). *A metaphysics for freedom*. Oxford University Press.

Strawson, G. (2000). The unhelpfulness of indeterminism. *Philosophy and*

Phenomenological Research, 60, 149–155. https://doi.org/10.2307/2653435

Van Inwagen, P. (2002). Free will remains a mystery. In R. Kane (Ed.), *The Oxford handbook of free will* (pp. 158–177). Oxford University Press.

Waller, B. (1988). Free will gone out of control. *Behaviorism, 16*, 149–162.

Watson, G. (1975). Free agency. *Journal of Philosophy, 72*, 205–220. https://doi.org/10.2307/2024703

Wolf, S. (1987). Sanity and the metaphysics of responsibility. In F. Schoeman (Ed.), *Responsibility, character, and emotions* (pp. 45–64). Cambridge University Press.

Wolf, S. (1990). *Freedom within reason*. Oxford University Press.