Better Will: A Critique of Sam Harris's *Free Will*

This essay is a critique of Sam Harris's book *Free Will* (2012) as well as his comments on free will more broadly. I think Harris often makes one-sided and misleading remarks on the topic, which I believe merit critique and correction.

To be clear, what I take issue with in Harris's view is not the assumption that human beings and human decisions are purely physical in nature. On the contrary, like Harris, I endorse a naturalistic worldview.

Similarly, my issue is not with the claim that we do not have free will in a sense that, as Harris writes, "cannot be made conceptually coherent" (p. 5). Indeed, I will not be arguing for the existence of "free will", nor will I entertain any discussion about what most people mean by "free will" (a term that tends to be poorly defined and which people often seem to understand in rather different ways; I think the term "free will" generally brings more confusion than clarity).

Instead, my issue with Harris's remarks on free will mostly has to do with his consistent emphasis on limiting perspectives — perspectives that are not a necessary feature of the view that we are purely physical in nature. As we will see, these limiting perspectives are taken to such an extreme that Harris often makes outright false claims that greatly understate our abilities and capacities.

Downplaying Our Knowledge

One of the main problems with Harris's book is his consistently downplaying what we can know. For example, when writing about two murderers, Harris writes (p. 4):

> Whatever their conscious motives, these men cannot know why they are as they are. Nor can we account for why we are not like them.

I think this is false. Murderers can surely know *something* about why they are as they are, and the rest of us can likewise know and explain a lot about why we are *not* murderers, from the level of our genes and upbringing to the level of our beliefs and values. In fact, there is a *lot* we can know at these respective levels.

Sure, we cannot know *everything* about why we did not end up as murderers, nor can we trace our explanations back indefinitely; there are limits as to what we can know about ourselves. Yet that

such limits exist is a rather trivial point. So the statement above seems to be either false (on a straightforward reading) or trivial (on a reading that would make it true).[1]

Downward Comparisons

Harris also tends to make what could be called downward comparisons, namely comparisons to those who seem to lack certain desirable capacities. For example, he makes the following point about the same two murderers (p. 4):

> As sickening as I find their behavior, I have to admit that if I were to trade places with one of these men, atom for atom, I would *be* him: There is no extra part of me that could decide to see the world differently or to resist the impulse to victimize other people.

One could make a similar point in the opposite direction. For example, one could draw a comparison to someone who is extremely morally virtuous, compassionate, and helpful for the world — a person with a high level of motivation to help other sentient beings; someone who continuously resists strong temptations to indulge in selfish desires and who instead deliberately pursues a deeper set of values, guided by a belief in the utility of personal agency and responsibility.

If Harris were to trade places with this person, he would *be* this person, and, importantly, in that case there *would* be a part of him that is able to resist lower impulses and desires (as is true to some degree of most people). He would be guided by deeper values and by a belief in the utility of personal agency.

Does it make any difference if we consistently make downward comparisons — as opposed to invoking a more balanced set of comparisons and examples — when discussing our agency and capacities? I think it does. For virtually any ability or capacity, we can identify a continuum from lower to greater levels of that ability. And if most people fall closer to the higher than the lower end

---

1 Perhaps Harris's point is that we cannot know or account for the *ultimate* cause of why we are as we are. Yet by this standard, it seems that we cannot account for or explain *anything*, since we arguably do not know the ultimate cause of anything's existence. In other words, this *ultimate* sense in which we cannot account for why we are as we are is a generic one that applies to *all* domains of knowledge. Maybe that is interesting as an epistemological matter (or maybe not), but it is not clear why it has particular relevance when it comes to accounting for how we are. In the standard sense of "explain" and "account for", the sense that we use in both everyday life and scientific discourse, we indeed can explain and account for (a lot about) why we are as we are and why we decide as we do, and there is hardly anything particularly special about these domains of knowledge compared to others.

of this continuum, it becomes quite misleading if we consistently focus on the lower end and imply that we are essentially all like that — especially when there are real and qualitative differences across this continuum.

By analogy, consider skill at the game of basketball: there is a wide continuum of human ability, and most people are far removed from the skill level of a newborn baby. After all, most people can pick up the ball and throw it, occasionally even making a basket.

Of course, one could highlight that if we were to trade places with a newborn baby, we would not be capable of picking up and throwing a basketball, and one could likewise highlight that everyone along this continuum of ability are physical and causally operating organisms to the same degree as newborns are. Yet the focus on newborns becomes misleading when we fail to acknowledge the real differences in capacities across this continuum: while a newborn does not have the ability to pick up the ball and throw it, most people *do*. Relative to newborns, most people do have "extra parts" — or rather more developed "parts" and skills — that enable them to pick up and throw a ball.

The same goes for the ability to resist impulses and to act on moral values. If we consistently draw comparisons to people who have the equivalent of a newborn's (lack of) self-control and moral capacities, and if we imply that we are all essentially like newborns, we risk downplaying the real moral capacities that most people *do* possess. More than that, we risk understating the degree to which we can further improve our moral capacities, just like most of us can improve our basketball skills if we acknowledge the potential and devote ourselves to it.

To be sure, this is not a matter of being non-mechanistic. On the contrary, it is a matter of developing the kinds of mechanistic tools that enable certain capacities, such as the capacity to catch and throw a ball, or the capacity to resist impulses and to identify and pursue the best path among a broad set of envisioned paths.

Downward Metaphors

Similar to his downward comparisons, Harris also consistently employs what could be called "downward metaphors". For example, when describing humans and human decision-making, he uses metaphors such as "clockwork", "biochemical puppet", and "neuronal weather patterns".

Of course, Harris's basic point is that our decision-making is purely physical and mechanistic. Yet a problem with each of these metaphors is that they are disanalogous to the mechanics of human decision-making in major ways, to the point of being bad metaphors. After all, there are many key elements to be found in human decision-making that we cannot find in weather patterns,

clockworks, or puppets. These include foresight, intentions, planning, learning, problem-solving, self-reflection, moral values, and so on. In short, the systems Harris invokes have zero agency and they are not in the least responsive to reasons and incentives.

A more accurate metaphor involving an obviously mechanical system would be a self-learning AI-driven robot. Such a system would at least possess more of the key elements listed above, and it would not be entirely shorn of foresight and agency.

Granted, the downward metaphors are more provocative and attention-grabbing. Yet it seems worth using more accurate metaphors when we can — or at least avoiding a one-sided reliance on only the crudest mechanical metaphors — again because a one-sided downward emphasis risks conveying a misleading picture of human capacities (just like a one-sided upward emphasis would risk conveying an unduly idealistic and overstated view of human capacities).

Making Our Wills

Harris writes (p. 5):

> Our wills are simply not of our own making. Thoughts and intentions emerge from background causes of which we are unaware and over which we exert no conscious control.

First, the claim that "thoughts and intentions emerge from background causes of which we are unaware" seems to be yet another case of downplaying our knowledge. After all, there is much we can and typically do know about the factors that drive our current thoughts and intentions, including our prior experiences, knowledge, and reflections.

Second, is it true that we cannot make our own wills in any real sense? For example, can we not, through reflection and effort, change and remake our wills to some degree? It seems that we can. Indeed, as Harris argues in the context of ethics, it is possible for us to care about and want the wrong things in life, and it seems possible to change what we care about and want in life based on deeper reflections about what matters. For some people, such reflections seem to have changed their wants and priorities profoundly.

So while our wills are not *ultimately* of our own making independently of prior causes, we can still reflect on and update our wants and values, and we can still make efforts to self-improve just the same. That is an important sense in which our wills *can* be of our own making — in the sense of partly being the result of our efforts and deliberations. And to the extent that we can improve our wills, it matters that we do so.

A Dilemma of Responsibility

I find this quote from Harris to be interesting (p. 5):

> Either our wills are determined by prior causes and we are not responsible for them, or they are the product of chance and we are not responsible for them.

This can sound plausible when stated in such general and simple terms. Yet would it be similarly convincing if we listed some more specific examples of how our wills are "determined by prior causes"? For example, we could specify it in the following terms:

> If our wills are determined by prior causes — such as by our prior reflections, attitudes, and deliberations — we are not responsible for them.

When we add these specifics, the claim that we are not responsible for our wills becomes less convincing. After all, it seems dubious to claim that we are not responsible in any real sense for that which results from our own attitudes and deliberations, in part because the everyday meaning of "we" typically includes things like our attitudes and deliberations.[2]

In this way, the more detailed framing helps expose how an abstract description like "determined by prior causes" can effectively hide some of the relevant processes that it is supposed to cover in a comprehensive way.

Acts of Volition Arising Spontaneously?

Harris writes the following about the subjective experience of volition (p. 6):

> Seeming acts of volition merely arise spontaneously (whether caused, uncaused, or probabilistically inclined, it makes no difference) and cannot be traced to a point of origin in our conscious minds.

This seems to downplay both our knowledge and the possible depths of human deliberation. Specifically, acts of volition are often not just arising spontaneously and without us being able to

---

2 There is also an ambiguity in that the word "responsible" has distinct meanings, such as (1) being the cause of something, and (2) being accountable for something. Yet we can be responsible to some extent in both these senses of the term. For example, if our current intentions are partly the product of our attitudes and deliberations, then we (if "we" are understood to include our attitudes and deliberations) are partly responsible for them in the causal sense, even though we are not their ultimate cause. Likewise, it seems plausible to maintain that we can be held accountable for our intentions if they are partly the product of our attitudes and deliberations.

trace (much of) their origin in our conscious minds. For example, in decisions that we have carefully deliberated over for a long time, our acts of volition may be the product of a vast array of considerations and long chains of reasoning of which we are quite aware — we might even have enumerated them in writing — and these considerations may give us a clear sense of why we made the decision we did.

Likewise, there are real senses in which our volitions sometimes can be traced to a point of origin in our conscious minds. For example, if a person experiences a state of intense suffering, and if the felt awfulness of this experience leads the person to decide to make the prevention of suffering their life's mission, it is plausible to claim that the conscious experience of suffering played a pivotal and initiating role in this decision. There would be a strong sense in which this decision to prevent suffering *could* be traced to a point of origin in that person's conscious mind, even though it is preceded by prior events.

Has Our Brain Already Determined What We Will Do?

Harris claims the following based on experiments in which scientists were able to partially predict simple human decisions via brain scans, even before people became aware of their decisions (p. 9):

> Some moments before you are aware of what you will do next—a time in which you subjectively appear to have complete freedom to behave however you please—your brain has already determined what you will do. You then become conscious of this "decision" and believe that you are in the process of making it.

The claim that your brain has "already determined what you will do" before you are aware of it seems far too strong. The word "determined" suggests that the initially planned action is unchangeable, yet if, for example, some unforeseen event happens in the external environment, we are generally able to change our initial plans quite rapidly, even if it happens without conscious awareness. So to say that our brains have already *determined* what we will do before we become aware of it understates the general adaptability of our planning and decision-making abilities. It seems more accurate to say that we form provisional intentions and tentative plans, which remain subject to change based on new information and feedback.

Furthermore, it is worth being clear that the pattern Harris describes — that our brain has already determined what we will do while we mistakenly believe that we are in the process of making a decision — often does not apply when we are in the process of making a decision. For example, when we are engaged in detailed deliberation that involves seeking out new information and

weighing many different reasons over a long period of time — say, a period of weeks or months — it is not the case that our brain has already determined what we will do at the outset of the decision process, and we are not mistaken to believe that we are in the process of making a decision during that long period of seeking new information and weighing our reasons. (To be clear, Harris does not claim otherwise, yet he does seem to mostly ignore these more deliberative and conscious decisions throughout his book — the kinds of decisions that do not conform to his description of us as being clueless about why we choose as we do.)

We should not conflate these elaborate decision processes with trivially simple ones. Again, there is a real and extremely broad continuum here, and if we consistently focus on the simple end of the continuum, we risk overlooking and neglecting the more advanced end. This may be a costly mistake, especially if our ability to create a better world depends on our engaging in advanced and elaborate decision processes.

Understating Agency

Harris writes (p. 9):

> I cannot decide what I will next think or intend until a thought or intention arises. What will my next mental state be? I do not know—it just happens.

These remarks emphasize passivity over agency, to an extent that again seems inaccurate. Is it really true that we cannot decide what we will next think until a thought eventually arises? Is it really true that our next thoughts and mental states "just happen"?

It seems that there is a strong sense in which these claims are not true, in that, for example, we can plan to sit down and systematically think about a given topic, thereby intentionally guiding what we will be thinking about. Sure, we cannot altogether settle what our thoughts will be in advance, but that would also render our process of thinking about and exploring a given topic rather pointless. We can still decide what we will think about and explore next, and we can base such decisions on considerations about what seems most important to think about. In other words, we need not passively wait until thoughts appear; we also have the ability to actively and deliberately pursue certain thoughts and questions.

The same applies to our future mental states: we can to some extent direct our mental states in desired directions, such as by engaging in meditation exercises or consuming particular substances that tend to produce certain mental states. In this sense, our mental states do not "just happen" without us having any degree of influence or control. (Again, Harris would hardly disagree with

these points, yet he nevertheless seems to greatly downplay our degree of agency and control in his statement above.)

Similarly, Harris writes (p. 19):

> I cannot determine my wants, or decide which will be effective, in advance. My mental life is simply given to me by the cosmos.

But as we saw earlier, there are important senses in which we *can* determine our future wants, or at least strongly influence them. For example, if we know that having a first drink leads to a stronger desire for a second drink, then we are in effect choosing to have less of that future desire by abstaining from the first drink.

Likewise, we can take steps to make some wants more effective than others, such as by deliberately strengthening some of our existing wants and motivations over others (e.g. we may decide to strengthen our compassionate motivations through compassion training, thereby potentially reshaping our wants and motivations quite substantially over time). In addition, we can change the incentives of our future selves. For example, we can set up systems such that we have to pay money if we follow our lower wants over our higher wants, which can help make our higher wants more effective. In these important ways, our wants, incentives, and mental lives are not "simply given to us by the cosmos" without any room for agency or control on our part.

Denying Agency

Harris goes on to seemingly deny human agency altogether (p. 25):

> There is no question that our attribution of agency can be gravely in error. I am arguing that it always is.

A problem with this claim is that the meaning of "agency" seems unspecified. And when making an extremely strong claim like "our attributions of agency are always gravely in error", it is worth being clear what "agency" means exactly. After all, by some fairly common definitions of "agency", Harris does not deny human agency. For example, Harris later writes the following (p. 38):

> Many people believe that human freedom consists in our ability to do what, upon reflection, we believe we should do—which often means overcoming our short-term desires and following our long-term goals or better judgment. This is certainly an ability that people possess, to a greater or lesser degree …

The ability that Harris here describes, and which people certainly possess, is essentially "agency" by a standard definition of the term: the ability to perform intentional actions (see e.g. Schlosser, 2019). Harris thus affirms "agency" by a common definition of the term, whereas he seems to employ an unusual definition of "agency" when he claims that all attributions of agency are gravely in error.

Harris also writes about agency in his book *The End of Faith* (2004). He writes (2004, p. 274):

> For our commonsense notions of agency to hold, our actions cannot be merely lawful products of our biology, our conditioning, or anything else that might lead others to predict them …

But again, is this claim correct? Do our commonsense notions of agency really require our actions to be the product of something beyond our biology and the information we have acquired in life? That is hardly the case if we take our biology and acquired information to include our intentions and values. Harris leaves out these key elements in his formulation, yet as we did earlier, we can include such more specific and more agency-related terms and see whether the revised claim seems plausible:

> For our commonsense notions of agency to hold, our actions cannot be merely lawful products of our intentions, our values, our deepest commitments, or anything else that might lead others to predict them.

In this formulation, the claim seems dubious. If others can predict our actions based on our values and our deepest commitments, then this seems fully aligned with commonsense notions of agency. And this holds true even if our values are altogether lawful. For example, our values might involve obeying the statutory law. Or they might involve always acting in line with some deontological moral law, or always acting so as to bring about the best consequences, or acting on some other guiding principle. Such broadly "lawful" ways of acting are not just compatible with our commonsense notions of agency; they arguably reflect the epitome of agency as commonly construed.

Tensions with the Possibilities of the Moral Landscape

In his book *The Moral Landscape* (2010), Harris defends a view of ethics centered on the well-being of conscious creatures. Specifically, among the possible decisions and actions we can take, Harris argues that we ought to pursue those that maximize the well-being of conscious creatures. He writes (2010, p. 40):

> Are there good and bad paths to take across this landscape of possibilities? Of course. In fact, there are, by definition, paths that lead to the worst misery and paths that lead to the greatest fulfillment possible …

There is a tension between this perspective and the perspective that Harris adopts when talking about free will. That is, in *The Moral Landscape*, Harris generally emphasizes a landscape of possibilities and alternative paths, while in *Free Will*, he writes that it is empty and meaningless to talk about "could have done otherwise" or being able to do otherwise (2012, pp. 39–40, 44).[3]

So it seems that Harris acknowledges that there are senses in which it is meaningful and legitimate to talk about possibilities and alternative paths — for example, possibilities and paths that are accessible in the world around us, and which we are able to pursue if we decide to pursue them.

Yet in *Free Will*, it appears as though he mostly overlooks those (in his own view) valid ways of speaking about possibilities and alternative paths. Moreover, he seems to overlook that this might be what people often have in mind when they talk about being able to do otherwise — namely nothing stronger or weaker than what Harris himself refers to when he talks about a landscape of possibilities and alternative paths, or indeed when he himself talks about what some people could have done otherwise in specific situations.[4]

Criminal Justice: Does Free Will Matter for Consequentialists?

Harris seems to argue that his view on free will has significant implications for criminal justice. For example, he writes (2012, p. 1):

> Without free will, sinners and criminals would be nothing more than poorly calibrated clockwork, and any conception of justice that emphasized punishing them (rather than deterring, rehabilitating, or merely containing them) would appear utterly incongruous.

---

3 Harris also touches on free will in *The Moral Landscape*, where his claims and emphases mirror those found in *Free Will*.

4 For example, in a recent interview, he talks about what he thinks presidential candidate Kamala Harris could and should have done otherwise during her campaign (The Bulwark Podcast, 2024). This suggests that Harris himself does not view all talk of "could have done otherwise" as meaningless, and that he indeed recognizes that there are perfectly reasonable and legitimate ways to talk about "could have done otherwise" and being able to do otherwise, in contrast to his unqualified statements in *Free Will* suggesting that all such talk is meaningless (2012, pp. 39–40, 44).

Similarly, he writes that if we acknowledge "the ultimate origins of human behavior", then "the logic of punishing people will come undone—unless we find that punishment is an essential component of deterrence or rehabilitation" (2012, p. 56).

Yet one can question whether Harris's view on free will — or a recognition of the ultimate origins of human behavior — would or should have any implications for criminal justice. For example, if one holds a consequentialist view of the kind that Harris endorses — a view aimed at creating the best outcomes for sentient beings — then the core commitment to creating the best outcomes seems unaffected by our stance on free will or the ultimate origins of human behavior.

That is, whether people have some ultimate free will or not, the commitment to creating the best outcomes would seem to stand just the same: from a consequentialist perspective, we should not create needless suffering either way.

Of course, one might argue that Harris's view on free will could count as a reason to favor a consequentialist view in the first place. I suppose it could, yet I would argue that it is not a particularly good reason. Specifically, I would argue that the overriding justification for a consequentialist focus on reducing severe suffering is the intrinsic badness of severe suffering itself, and this justification is orthogonal to our views on free will (Vinding, 2020).

Conclusion

As Harris acknowledges, there is a real sense in which we have possibilities and alternative paths available to us. And a standard notion of agency is simply to pursue certain such possibilities and paths over others — something that Harris himself acknowledges we can do, and something he even argues that we morally ought to do. In this sense, agency is not an illusion, but is very real indeed. We are able to, and have good reason to, pursue the best paths and possibilities that we can envision.

Relatedly, from an agentic and forward-looking perspective, it may be worth shifting our focus from "free" will toward "better" will. After all, what matters most is that we embody and pursue a better will — ideally the best possible will — not that we have a "free" will. Indeed, if having a "free" will made the world a much worse place for everyone, we would hardly want it, and we would be right not to want it.

Conversely, if the best possible will made us and everyone else as well off as we could be, it would be desirable to cultivate and act on this best possible will, regardless of how "free" or "unfree" it might be. (That being said, cultivating and pursuing a "better will" likely involves many of the elements that tend to be associated with "free will", such as deep reflection, carefully weighing different options, and a strong sense of resolve and direction.)

In short, what is ultimately worth wanting and acting on is a *better* will, and hence it matters that we strive to develop and act on a better will.

Bibliography

The Bulwark Podcast. (2024). Sam Harris: Our Democracy Is Already Unraveling. https://audioboom.com/posts/8611883-sam-harris-our-democracy-is-already-unraveling

Harris, S. (2004). *The End of Faith: Religion, Terror, and the Future of Reason*. W. W. Norton.

Harris, S. (2010). *The Moral Landscape: How Science Can Determine Human Values*. Free Press.

Harris, S. (2012). *Free Will*. Free Press.

Schlosser, M. (2019). Agency. https://plato.stanford.edu/archives/win2019/entries/agency/

Vinding, M. (2020). *Suffering-Focused Ethics: Defense and Implications*. Ratio Ethica. https://magnusvinding.com/wp-content/uploads/2020/05/suffering-focused-ethics.pdf