Teorema, 2025, 44(1), pp. 1-10
https://doi.org/10.30827/trif.34395

In defense of our lives as biological machines

Robert Sapolsky

Stanford University

Abstract: The preceding pieces thoughtfully argue that we possess free will, both of the type that we would want in the moment, and of the type that has determined the sort of person we turned out to be. Moreover, they argue that this overwhelmingly fits our everyday intuition that we can be free at important moments, and that such moments can reflect our ability to consciously choose to amplify or negate the effects of circumstance upon us. In this piece, I heartily and respectfully disagree with all these points.

Keywords: *free will, biological determinism, moral responsibility, neuroplasticity, human behavior*.

Resumen: Los textos anteriores argumentan de manera reflexiva que poseemos libre albedrío, tanto del tipo que desearíamos tener en el momento presente como del tipo que ha determinado la clase de persona en la que nos hemos convertido. Además, sostienen que esto concuerda plenamente con nuestra intuición cotidiana de que podemos ser libres en momentos importantes, y que tales momentos reflejan nuestra capacidad consciente de elegir amplificar o contrarrestar los efectos que las circunstancias tienen sobre nosotros. En este texto, discrepo de todo ello de forma respetuosa pero decidida.

Palabras clave: *libre albedrío, determinismo biológico, responsabilidad moral, neuroplasticidad, comportamiento humano*.

Received: 4 June 2025 **Accepted**: 4 July 2025 **Published**: 30 July 2025

Teorema. Revista Internacional de Filosofía
ISSN/ISSN-e: 1888-1254

* This work is part of Book symposium on Robert Sapolsky's *Determined: A Science of Life Without Free Will*, organized by Guest Editor: Jesús Zamora Bonilla. The symposium is by invitation only, but all published papers undergo extensive peer review.

I am delighted to discuss free will with these five scholars and warmly thank Jesus Zamora Bonilla for the opportunity to do so. Rather than responding to commentaries in sequence, I organize my response by themes.

1. Intuitionism

A philosopher and a biologist walk into a bar and, before the door even closes behind them, discover that they have different definitions of "bar," "into," "walk," and, to the biologist's horror, demonstrate that while the biologist can't quite articulate a definition of "a," the philosopher has a dozen handy. This is the reason why any conference on an interdisciplinary topic devotes the first day to definitions. Thus, it is not surprising that these papers consider what is meant by our will being free (with Mark Balaguer offering a particularly detailed dissection of the differing definitions); I thank all that this did not generate what Zamora called a menacing "semantic forest."

Amid the resulting complexities, Andrew Vonasch, Scott Danielson and Alfred Mele turn to the intriguing new field of experimental philosophy which explore the "ordinary usage" of the concept of free will. In their study, subjects were asked to decide if free will was occurring, presented with a clever progression of scenarios that varied the immediacy, explicitness and implacability of the biological and psychological influences on behavior. The results showed that a sizeable majority of people saw free will persisting despite the influences of genes, neurons, cultural norms and our pre-existing tastes. As an important negative control, a fifth scenario invoked Dennett's "nefarious neurosurgeon," where a futuristic drug and brain implant control someone's behavior. Tellingly, most participants saw no free will in this circumstance.

There is, of course, a circularity in the fifth scenario, where the honest person in the scenario is forced to do something unethical and "irresistible"; in that circumstance, definitionally, resisting the irresistible is futile. Moreover, amid it being fascinating that 79 – 92% of people perceived free will in the first four scenarios, it is equally fascinating that 8 – 21% did not. Nonetheless, the important conclusion from this study is that while most people see free will as withstanding various biological and psychological factors, they believe that some circumstances preclude free will.

The obvious problem is that folk philosophy and intuitionism are poor metrics to use as litmus tests. For example, the average test subject's intuition as to whether it is okay to sacrifice an innocent person to a runaway trolley in order to save five other lives varies dramatically, depending on whether the test subject or an autonomous robot will push the victim onto the tracks. The average test subject's intuition as to whether it is okay to enslave some types of people varies dramatically, depending on whether the researcher is carrying out the study in the 21st or 17th century. Intuition about agency, responsibility and culpability are moving targets, subject to the whims of place, time, wording and the flapping of butterfly wings.

In other words, while fascinating, folk philosophy is of limited use for this discussion. In much the same way, my own lay intuitions are probably not useful when it comes to understanding why a carburetor, an economy, or someone's spirit is "broken." As such, it is questionable for Vonasch et al. to conclude that intuitive definitions of free will given by most people "should" be embraced and "properly assigned" prima facie.

2. The moment of choice and the difference between now and what led up to now

Fortunately, the commentators and I agree that the absence of free will in a contemporary sense does not mean that the future was already determined right after the Big Bang -- the spirit of Laplace's Demon. Such predictability is not possible because of the non-linearity and chaoticism of complicated things (e.g., the universe…). Thus, any given moment contains multiple possible futures destined to be shaped by our actions. In that regard, I emphatically agree with Balaguer that he can be Hume-free and act on his desires in choosing chocolate over vanilla ice cream. As can the rest of us.

My key point in this section harks back to the challenges of definition, because at many points, a number of the commentators and I are focusing on different free wills. There is being Hume-free in the senses just outlined, with the seeming capacity to a) consciously choose something, b) have a reasonably accurate prediction of what the outcome of that choice will be, and c) know that there was no coercion. This is true for most of us much of the time and fits the layperson's conventional sense of a free will worth wanting; it is (usually?) good if we can get chocolate ice cream when we desire it.

However, I feel that this is not the place to look when it concerns the version of free will with which our lives are judged. As noted by Kevin Mitchell, this takes us to Schopenhauer and his truism "man can do what he wants, but not want what he wants." Similarly, we cannot will ourselves more willpower than we possess or choose what we will think next. We may be free to choose chocolate, but we were not free in having become someone who would *want* chocolate in that place and time. To combine Zamora and Sartre, we are condemned to choose, but not freely.

This raises three questions: What is more important in assessing how "free" you are – the intent you form or your pursuit of that intent? How did each of us became the "sort of person" that we are? Is there really no free will in this domain?

What is more important in assessing how "free" you are – the intent you form or your pursuit of that intent?

When it comes to making sense of how the world works and its sources of injustice and misery, I think the answer is clear. The freedom to act on our intentions, while important, pales in comparison to whether life's circumstances outside our control have filled us with the intentions of an anarchist or Jacobite, a secular saint or demon, of someone capable or incapable of love. The ability to pick chocolate, the proximal sense of freedom is precisely what fuels our everyday perception of agency, where intuitionism most magnetically and myopically leads us to mistake a sense of agency for agency itself. And importantly, the emphasis on its importance also reflects it being a more tractable realm to try to understand, whether life's circumstances have turned you into a philosopher or biologist.

How did each of us became the "sort of person" that we are?

You walk into the ice cream store and request chocolate, *the thing you desire.* But that is not the only behavior available to you. You could choose a different flavor, rob the store, propose marriage to the stranger working there or quack like a duck. How

did life's circumstances come to turn you into the sort of person (and as an important point) *in that moment*, whose nervous system over the scale of milliseconds generated the utterance of the word "chocolate"?

Your immediate sensory and social environment mattered (and thus my emphasis on "in that moment" – events in the previous minutes play a role in making you the sort of person you become in that subsequent minute). Maybe you'd normally prefer vanilla but the aroma of chocolate made you into a different sort of person thirty seconds later. Maybe you note that the vanilla ice cream looks moldy, or that the chocolate is both free and fat-free. Those prior minutes matter for who you are in that moment.

But the prior hours to days do as well. Maybe you feel like you've been in a rut and plan to order something different from your usual chocolate, but it's been a miserable day and your elevated levels of glucocorticoid stress hormones make you more risk averse (which they do); thus, you feel unable to withstand the potential disappointment of a new flavor not tasting as good as chocolate. Maybe because of your elevated levels of androgenic hormones, you interpret the neutral facial expression of the person working there as being hostile (an effect that androgens have), and you storm out. Maybe because of where you are in your ovarian endocrine cycle, your nose is atypically sensitive to the faint smell of phytoestrogens in the soy ice cream. Those prior hours to days also matter for who you are in that moment.

As do the prior years to decades. Maybe because of the trauma that has left you with social anxiety, you cannot enter the store. Maybe because of your knowledge that the people who make the chocolate ice cream are co-religionists, you choose chocolate even though their ice cream is bland. Maybe life has privileged you with unexpected pleasures and you feel that the next new thing may become your favorite, and you pick the beef tripe ice cream. The point here is not just that long-term experience matters, but long-term experience has *changed your brain*, the realm of neuroplasticity. Some brain regions have expanded or shrunk, some networks of neuronal projections have grown more or less complex, some brain-specific genes have been permanently activated or silence via epigenetics. There is a nuts-and-bolts neurobiology as to what "the years" have done to you.

Similar mechanisms will explain how your period of greatest brain development – from fetal life to late adolescence – will have, for example, sculpted your frontal cortex such that you can't resist ice cream despite being on a diet. And further back, there are your genes that code for what taste receptors you have in your tongue. And whether your ancestors were hunter-gatherers without domesticated animals and thus bequeathed you with a culture in which dairy is solely for infants. And whether evolution made you a member of a species that will forgo ice cream and instead predate the person working there.

The critical point is "the sort of person you turned out to be" is entirely a function of what came before, and what came before is entirely a function of the biology over which you had no control and its interactions with environment over which you had control. This is the science of why we can pursue what we desire but cannot choose what we desire.

Influences versus determinants: are we truly "entirely" a function of our history of biology x environment?

Thus, my rejection of free will is based on the claim that all we are is our uncontrolled history. Mitchell notes a circularity here if one concludes that we don't have free will now because we never had free will. But there is a complementary circularity as well in the idea that we are exercising free will when we make a choice because we chose to become the sort of person who values that choice. Both seem equally challenging infinite regresses to me.

Mitchell frames my emphasis on our uncontrolled histories mechanistically, pointing out that my view is that "the brain is pre-configured in such a way as to give different "weights" to different kinds of signals or information (representing beliefs, goals, desires, etc.)." But Mitchell and some of the other commentators question whether such pre-configuration determines us or merely *influences* us. This would be because "we can't pre-state all the relevant first- and second- and third-order weights because the space of possible combinations across all scenarios we might encounter is effectively infinite and unknowable in advance. This kind of combinatorial explosion makes the problem computationally intractable."

I believe that because of our lack of free will, we are all machines. But that metaphor potentially leads someone, I've now discovered, towards the wrong kind of machine. We are not examples of "more is different" applied to a microwave; that would indeed require us to be automatons, fruitlessly tackling infinity with a point-for-point approach. Instead, we are examples of "more is different" being applied to machines like paramecia, sea slugs and chimps. Both they and we bypass the computational intractability of having to code for specific pre-configured responses to an infinity of possible eventualities. Instead, what we do is generalize and categorize. And we humans excel at this (something explored in Simon Baron-Cohen's *The Pattern Seekers*).

We generate pre-configured categories of context-dependency on a behavioral level – "Don't shove frail old people out of the way…unless something in the category of out-of-control motor vehicles is hurtling at them." Our bodies do the same on a physiological level – "Testosterone does not generate aggression…unless the organism is experiencing the category of behaviors that constitute threats to his social status." Likewise on the genetic level – "having the short-s allele of the 5HTTLPR serotonin transporter gene does not increase the risk of depression…unless the individual experienced examples of what we categorize as significant trauma in childhood."[1]

Of course, our dealing with infinite possibilities by applying categories to achieve Vervaeke's (2012) "relevance realization" opens a can of worms because, as emphasized by Mitchell, we have brains that value efficiency over precision. This is shown by the fallibility of our heuristic shortcuts, the world of behavioral economics pioneered by Tversky and Kahneman (1979). It is logically equivalent to cure all the cases of one disease or to cure half the cases of two diseases – but the former usually feels more satisfying. It is not possible for part of a set to be bigger than the set itself, but we often endorse that if the former is described in more detail than the latter. A prize is of equivalent value whether you have received it and now must give it back or whether you never got it in the first place – but we loathe the former. These heuristic shortcuts readily lead us astray.

[1] For aficionados, the replicability of this landmark finding (Caspi et al., 2003) has generated great controversy in psychiatric genetics. For what it's worth, my personal bias is that the finding is generally sound.

Another problem is that the infinite number of circumstances that life presents can be categorized in an even bigger infinity of possible ways; how did you become the sort of person who forms the categories that you do and fill them with their constituent parts? After all, this will determine what you consider to be good art or dastardly acts. And these play out on both the conscious and implicit levels. This is why a police officer with an implicit bias against, say, leprechauns, looking in a crowd to spot who likely just committed some crime, will unconsciously home in on leprechaun faces. Like sea slugs (but not like microwaves), we biological machines compress the infinite into categories whose uncontrolled boundaries, salience and mutability comprise the "sort of person" circumstance has made us. Insofar as all of infinity can be compressed in this way, the factors that create our categories generate far more than mere influences.[2]

3. Being torn

This previous section argues that the sum of biology x environment generates causation rather than influence. One of the commentaries, however, explores the possibility that our meaningful freedom comes from circumstances where there are not even influences.

Balaguer posits that "there will often be a range of possible actions with indistinguishable predicted utility." These are circumstances of *torn decisions*, where two options consciously are equally desirable; if that feeling remains at the time of choosing between them, that decision is free of prior influences, "*is* the fork event," a "non-decisional action" (his emphasis). It is here that libertarian free will dwells.

Appropriately, Balaguer emphasizes that true torn decisions are rare because of the definitional emphasis on the non-decisional decision being entirely conscious. He readily accepts that many, most, even nearly all decisions are not torn because of determinants that we are not aware of, "by our subconscious beliefs and desires, or by magnetic stimulations to the brain, or by subliminal advertising, or whatever." I think we'd both agree that even a single instance of a Platonically pure torn decision would prove the existence of free will. The difference, of course, is that I think that can *never* occur in we biological organisms.

Why? The recent centuries in which science contributed to the free will debate consists of people repeatedly realizing that "I had no idea that biology had something to do with Behavior X." And thus, we have come to learn that adverse perinatal events explain schizophrenia more accurately than do myths of malignant mothering. That cytoarchitectural anomalies in the cortex explain dyslexia more correctly than do charges of laziness. That frontocortical damage explains some criminal acts more meaningfully than do discussions of soiled souls. And with each passing day, science uncovers more of these subterranean biological factors.

Yes, yes, a response can be pointing out that biology x environment does not

[2] By the way, Balaguer frames my stance regarding influences a bit more broadly than I think is accurate. "A person is *Sapolsky-free* just in case at least some of their decisions have the following trait: which option was selected in the decision wasn't causally influenced by any events in the history of the universe." I don't actually subscribe to "in the history of the universe;" just the parts pertinent to how you became you in that moment. Thus, I am not claiming that an inability to fulfill your desire for chocolate ice cream represents a lack of free will if the history of the universe has led to the store being closed.

currently explain *everything* (and, when factoring in chaoticism, that it never will). A stance like Balaguer's suggests that free will is what we call ignorance. In this context, free will skeptics are often challenged to prove that there is no free will, an absence of proof/proof of absence conundrum. I believe that the explanatory matrix of insights regarding our behaviors – ranging from the neurochemistry of one millisecond ago to the evolutionary biology of a million years ago – is so rich that the tables need to be turned: prove where free will fits into this, explain how the movement of a muscle can itself be an uncaused fork event. It cannot be explained by quantum indeterminacy.[3] And it cannot be explained by random neuronal activity because it turns out not to be all that random;[4] more importantly, because just as it is problematic if the supposed free will that comprises our moral compasses is based on our ignorance, so too if it is based on randomness.

4. Change

Gloria Andrada focuses on a key issue in free will debates, namely change. Fortunately, all the commentators avoid the frustrating misassumption that rejecting free will means rejecting the possibility of change. It obviously occurs. People make a living removing tattoos, when it turns out that a client's tattoo proclaims a love that turned out not to last forever.

The question becomes how change occurs if there is no free will. And the non-simple answer is that rather than *choosing* to change, we are *changed* by circumstance, and as a function of who we turned out to be at the time we experience a circumstance. An example:

Three individuals go to see a movie about an inspirational topic (e.g. the 2004 film, Hotel Rwanda, about Paul Rusesabagina, a man who took unimaginable personal risks to save more than 1,500 people during the Rwandan genocide). And the three moviegoers emerge changed by the experience. One will forever be moved by Rusesabagina's story; one will forever admire the stirring cinematography; one will forever be irritated by how the theater was noisy and hot. It seems obvious that those three different responses reflect the prior circumstances that created three different people sitting down to watch the movie.

Moreover, being changed by an experience in a history-dependent way is readily extended to those three individuals then changing *other* people. The first person informs their friends about the moving story; the second about the artistry of the cinematographer; the third about how they should avoid that theater.

But Andrada focuses on a more interesting and subtle level of how change is compatible with a lack of free will. Like every organism, we are biological machines. However, crucially, we are the only organisms that can *know* we are machines and learn about the buttons that control them. We have meta-cognition. This leads Andrada to

[3] A conclusion that Balaguer and I share.

[4] When one examines the mechanisms underlying random neuronal firing and its regulation, it becomes clear that it is most accurate to state that there are occasions when the nervous system determines that it is a good time for some indeterminism. Thus, seeing free will in neuronal "randomness" becomes like claiming that it is free will when a theater student is given the parameters of an improvisation by the teacher in an acting class.

write, "can we, by interacting with and/or modifying this environment, transform or alter our own conditioning?" Absolutely – but without having to invoke free will.

Recall the first individual, who had left the movie theater greatly moved by the story. In a moment of meta-cognition, they think that they would benefit from learning more about the Rwandan genocide and vow to read books on the subject. Thus, as a result of their meta-cognitive decision, they will change further, becoming more informed about the subject. They will have altered their own attributes.

But this needs to be unpacked further in asking the same vital question raised above – how did they become the sort of person who would reach this meta-cognitive decision? Why do they happen to respect knowledge and reading? Did prior adversity make reading a way to cope with emotional distress? Did tenacious reading help them escape their family's poverty? Was reading a passion of the first person they fell in love with? All were different pathways to the construction of a particular type of brain. And in the same spirit, how did reading become their response to being moved rather than, say, volunteering to help with refugee resettlement?

This raises another level of questions. Will this person actually read those books, read half of the first one and lose interest, or never get around to it? Did they become the sort of person with sufficient frontocortical function to follow through on their goal?

Remarkable insight into this comes from the work of psychologist George Ainslee in his 1974 study Impulse control in pigeons (J Experimental Analysis Behavior 21, 485), highly influential research cited more than 1,250 times in the literature. Pigeons were given an equivalent of the famed "marshmallow test" of gratification postponement in children. In it, if they pecked at a lever, they would receive a food reward; however, if they resisted pecking the lever for a length of time, they would receive a larger food reward. Naturally, 95% of pigeons showed no self-control and pecked the lever.

Then, as the key elaboration, pigeons were now trained on two levers. The first was as before, yielding a small reward in response to pecking. The second, when pecked, *prevented* the first lever from being pecked, forcing the pigeon to wait long enough to get the larger reward. And pigeons pecked this second lever. Thus, a pigeon can both fail to show gratification postponement *and* choose to be forced into a situation where that failure does not matter. One does not have to anthropomorphize a type of consciousness that would lead a pigeon to think, "I know that my self-control is terrible so I'd better pecked the second lever." More efficacious use of our biological buttons does not require consciousness. Even a pigeon can ask to be tied to the mast when sailing past Sirens.

This is paralleled more explicitly in the realm of addiction. One recovering alcoholic can understand their buttons enough to know that they can go to a bar with friends and drink only soda water. One can know that this will not be successful and avoids entering the bar. One knows that their self-regulatory limits are one step further and avoid walking down a street with a bar.

The Ainslee study had one additional finding that cements its relevance to us – only 30% of the pigeons had the self-discipline to press the second lever that forced abstinence. And thus, whether considering pigeons or humans, the question becomes,

How did uncontrolled circumstance (biology x environment) generate a frontal cortex capable of pressing the machine's buttons effectively?[5] Thus, the lifetime of factors over which we lacked control determine whether we can learn the lessons of an experience, imagine a future that reinforces those lessons, and have the self-discipline to implement that imagined future. Change in all these variants is compatible with an absence of free will.

5. Conclusions [6]

We make conscious choices each day, knowing their likely outcomes, knowing that our choices are not the only ones available to a human in that circumstance; this is where we most strongly intuit the presence of free will. But although we may choose to act on a desire, we do not choose freely, because we are unable to choose *what* we desire.

This is because who we are at that moment is solely the outcome of the circumstances that made us. Those circumstances range from sensory stimuli in our immediate environment to this morning's hormone levels; from decades of neuroplasticity to the nature of our childhoods; from our fetal environment and genes to our legacies of culture and evolution. In theory, prior history can generate influences rather than determinism; in such a circumstance, we would choose between options that we consciously view as being infinitesimally equal; thus, in effect, the choice we would make would constitute its own history. But such infinitesimally equal choices are an impossibility because of the vast subterranean biological forces beneath the surface of our being. As a result, we are the sum of the biology over which we had no control and its interactions with environment over which we had control. Importantly, this does not preclude the possibility of change; like every organism, we are changed by circumstance as a function of who we turned out to be at the time of the circumstance. Crucially, however, we are the only biological machines that *know* that we are biological machines and learn the workings of its buttons. Moreover, circumstance has made some of us capable of accessing those buttons, to amplify or negate the effects of what chance presents us with. Nonetheless, biological machines we are.

References

Baron-Cohen, S. (2020). The Pattern Seekers: How Autism Drives Human Invention. Basic Books.

Caspi, A., Sugden, K., Moffitt, T., et al. (2003). Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science,* 301, 386-389.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47. 163-291.

[5] While pigeons do not possess the frontal cortex characteristic of mammals, they have a rudimentary equivalent.

[6] Amid this piece constituting a response to the views of Vonasch et al., Balaguer, Mitchell and Andrada that differ dramatically from my own, I've devoted relatively little space to Zamora's thoughts on this subject. This is because we are seemingly in agreement concerning all the substantial elements of the free will debate, and my main response to his piece is great respect for his expressing these ideas so much more clearly than I have.

Sapolsky, R. (2023). Determined: A Science of Life Without Free Will. Penguin.

Vervaeke, J., Lillicrap, T. P., & Richards, B. A. (2012). Relevance realization and the emerging framework in cognitive science. *Journal of Logic and Computation*, *22*(1), 79-99.