**Undetermined: Free will in real time and through time**

Kevin J. Mitchell

Institutes of Genetics and Neuroscience, Trinity College Dublin, Dublin 2, Ireland.
kevin.mitchell@tcd.ie

In his book "Determined", Robert Sapolsky (RS) argues that humans have no free will, and, indeed, that we have "no shred of agency" at all (Sapolsky, 2023). His argument is an interesting one, as it does not depend on the usual appeals to physical pre-determinism and causal reductionism, where all the real causes of behavior are presumed to reside at the lowest levels of physical systems, the dynamics of which are presumed to be deterministic.

RS seems to accept that the most fundamental levels of physical reality – those described by quantum physics – are genuinely indeterministic. He argues, however, that individual random quantum fluctuations or events will not manifest at macroscopic levels, because there are so many of them going on at any time that they will effectively wash each other out. Or, in what may be a slightly different argument, that higher-order thermal fluctuations in the brain will swamp out lower-level randomness:

"*People in this business view the brain not only as "noisy" in this sense but also as "warm" and "wet," the messy sort of living environment that biases against quantum effects persisting*." (page 221)

It's not clear where, in this picture, *the higher-level randomness*, or background noisiness, is supposed to come from, but actually, for the purposes of the discussion of free will, it's sufficient that it exists. The important point is that physical pre-determinism does not hold, *at any level*. This means the evolution of the system through time should not be completely pre-determined, but open – i.e., based on low-level physics alone, in most of the kinds of systems we care about, many things could happen (Del Santo, 2021; Del Santo and Gisin, 2019; Ellis, 2016; Potter et al., 2025; Smolin and Verde, 2021; van Strien, 2021).

Based on this position alone, it would seem that at the level we do care about – the level of neural goings-on – RS should endorse the view that, given some "input" to the system, many possible "outputs" could arise. However, he insists that, in any given scenario we might encounter, our (neurally instantiated) processes of decision-making and action selection will always lead us, inevitably, to a single outcome. He is thus making an unusual claim: that our behavior is determined, not *because* the underlying neural and physical substrates are

deterministic (which would locate all possible causality at those lower levels), but *in spite of* genuine indeterminacy at these levels.

This amounts to a form of macroscopic, or even mental causation of a kind that RS seems to otherwise wish to deny. After all, if the underlying neural components and processes are noisy (which they empirically have indeed been found to be (e.g., Faisal et al., 2005; Glimcher, 2005; Rusakov et al., 2020; Sanborn et al., 2024)), then the only way for outcomes to be determined, at the level of our behaviors, would be through some higher-order constraints to be at play. What RS seems to be endorsing is thus a kind of top-down *cognitive determinism*.

He seems willing to accept that we genuinely act for our reasons – i.e., that we have sets of beliefs and desires and other cognitive states, and that it is the "content" of these states that informs what we will do. That seems consistent with views that neural systems evolved as control systems – precisely to enable organisms to do things *for organism-level reasons* (Cisek, 2019; Mitchell, 2023). However, RS takes this idea one step further. He argues that the set of beliefs, desires, and other cognitive states that obtain in our minds at any moment are jointly sufficient to not just inform but *completely determine* what we will do at any moment, given any new set of sensory inputs. In a sense, we have *so much control* that we end up with no choice!

This is a bold move, given that there is no actual evidence for it. RS does cite lots of statistical evidence for the individual *influences* of all kinds of antecedent causes on our *patterns of behavior*. These include human evolution, our own genetics, the ways our individual brains happened to develop, in utero and early childhood exposures, and the traces of prior experiences. But no one disputes these influences – they are the things that collectively make us who we are (Mitchell, 2018). What is disputable is the idea that these factors collectively make us automata.

One gets a sense from reading Determined that RS sees the processes of decision-making and behavioral control as essentially passive and algorithmic – the workings of a great big stimulus-response machine. His view seems to be 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.). Though it's obviously enormously complicated, working through the algorithm is then a matter of taking some new inputs and simply waiting for the machine to spit out the answer. In the literature on the philosophy of action, this raises what is known as the "disappearing agent" problem (Pereboom, 2014). If our behavior in any scenario is determined by the collection of internal mental

(or neural) states at any moment, then what is there left for *the agent* itself – as a holistic entity – to do?

## The need for judgment

The problem with this model is the assumption that the weights given to the various parameters of any given situation – actually of *every* possible situation – could in fact be pre-set in this way. This is akin to thinking that the constitution of a country, along with its set of laws, and body of legal precedent would be sufficient to algorithmically determine the right sentence for anyone found guilty of a crime. We could just plug all the information into the Justice Algorithm, which would spit out the right answer, leaving nothing for judges to do. That is, no *judgement* would be required.

The problems with this idea should be obvious. We can't just run any set of new information through a totally predefined cognitive algorithm, because the parameters of the algorithm are always massively context-dependent. And 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 (Bossaerts et al., 2019; Rich et al., 2020).

That's not to say that a lot of our behaviour isn't fairly automatic. In many familiar or simple scenarios, we know what to do based on a simple set of habits and heuristics – no thinking required (Kahneman, 2011). But once things get more novel, and more complex, we can't just submit the information to a preconfigured algorithm. We have to figure out *how the algorithm should be configured*. We have to work out how to weigh up all the various factors, on the fly, in real time (Lemos, 2021). In other words, we have to do exactly what judges do – make a judgment!

Not just in real time, either – we have to do it *in good time*. And we don't have infinite compute or unlimited energy – we just have our limited brain, selected for efficiency, not precision (Sterling and Laughlin, 2015). Indeed, in many cases, it's not like there actually is *a right answer*, waiting to be found. Given the limited information (and considerable uncertainty) that an individual has at any moment, and all the potentially conflicting goals over which they are trying to optimise, simultaneously, there will often be a range of possible actions with indistinguishable predicted utility (Gigerenzer and Gaissmaier, 2011; Glimcher, 2004; Kahneman and Tversky, 1979). This is the well-known principle of "bounded rationality" (Simon, 1990). We can't always reach a definitive solution

– sometimes we just have to pick one without knowing in advance if it will prove to have been the best choice.

To navigate through a world that is constantly throwing up novel scenarios, we need to do work to judge the relative salience of different factors, relative to our suite of current goals. We need to infer not just what is out in the world, but which elements matter to us – what we should care about and pay most attention to and weight most heavily, given our current state, our ongoing projects, whatever behavioral agendas we are pursuing, and so on. In short, we need to make decision-making computationally tractable. John Vervaeke and various colleagues have called this process "relevance realization" and have presented some biologically plausible ways in which it can be achieved, allowing organisms to make their way pragmatically in an open-ended world. (Jaeger et al., 2024; Vervaeke et al., 2012; and see: Cisek and Kalaska, 2010; Mitchell, 2023).

This will inevitably involve lots of heuristics, rough estimates, and other means of *satisficing* – that is, trying to satisfy the myriad demands on behaviour, based on all kinds of competing considerations, in a reasonable amount of time with limited resources (Artinger et al., 2022; Simon, 1956). This doesn't *just happen* – it's hard work; it's something the organism has to do. (Cisek and Kalaska, 2010; Redish, 2013; Shadlen and Kiani 2013). It is definitively not just a great big stimulus-response machine passively churning through the steps of a deterministic algorithm. In fact, the only way that these cognitive-level processes could be deterministic is if the underlying neural and physical processes were completely deterministic. And we know (and RS seems to accept) that they're not.

If the neural computations are causally sensitive to semantic content, rather than detailed syntax, and those semantics relate to person-level concepts, and all that information is integrated in a hugely contextually interdependent way, and is used to direct behavior over nested timescales, in ways that cannot be either algorithmically or physically pre-specified, based on criteria configured into the circuits derived from learning, which embody reasons of the person and not any of their parts, then I would say that ***just is the person deciding what to do***. (Where the person – or any organism – is not a machine with decomposable parts, but an integrated self with continuity through time (Mitchell, 2023; Nicholson, 2013)).

Moreover, human beings in particular have the remarkable and perhaps unique capacity for metacognition, meta-reasoning, and metavolition (Ackerman and Thompson, 2017; Fleming et al., 2012; Fletcher and Carruthers, 2012). We can reason about our reasons, in real time, and change our minds in the process of

deliberation. Again, RS does not deny this – he refers, for example, to the prefrontal cortex as the bit of the brain that lets you "do the right thing, when it's the harder thing to do". Given that RS seems to accept the existence of these elaborate and sophisticated capacities for introspective, rational, reflective, real-time control of our own behavior, it's worth asking: "what else would be needed for RS to call this "free will""?

The answer seems to be: ***ultimate moral responsibility***. This is a notable shift from discussions that focus on the narrower (and frankly better defined) question of whether we have the capacity to act for our reasons. But it aligns with a long tradition of entangling questions of free will and moral responsibility. For some, a freely willed action just is – by definition – one for which you can be held morally responsible. So, why might RS not take the evidence that we really can do things for our reasons as sufficient to ground moral responsibility? There are at least three reasons.

First, he seems to take every advance in neuroscience, which reveals more of the mechanisms of decision-making, as diminishing any possible role that *you* could be playing. This simply rests on a dualist framing, where your brain is seen as making the decision, rather than seeing it as you (as an agent) *using your brain to make a decision*. There is no reason to accept this framing, as it rules out any kind of naturalistic free will from the get-go – only some supernatural force would satisfy.

Second, he might say that if you are not free *from any prior cause whatsoever*, when you make a decision, then you cannot be held responsible for it. This absolutist framing (which is fairly widespread) is also incoherent, as it is precisely continuity through time (i.e., the propagation of prior causes) that characterises life itself (Mitchell, 2023). The concern rests on the mistaken idea that prior causes must always be comprehensively necessitating, when in fact many factors may be *contributing* to any decision, while still leaving a decision to be made.

And third, there is the concern that those prior causes include things that have shaped our reasons in ways that we ourselves did not choose, and for which we should thus not be held accountable. This seems to be the real crux of the matter for RS. His view harkens back to a long tradition of arguments, captured by Schopenhauer in his phrase, translated roughly as: "man can do what he wants, but not want what he wants" (Schopenhauer, 1960). As described above, this reflects the notion that our intentions at any moment *simply arise* as a consequence of our current motivations and beliefs. That is, that the way our brain is currently configured leaves us with no choice at any moment – nothing that could be said to be *up to us*. The important point (for RS) is that because we

never had any choice in the past, that means we cannot be held responsible for any aspect of our current configuration – for our having gotten to be the way we are – and thus *cannot be held responsible* for the actions that issue from that configuration.

Note the circularity of this argument: we never have free will (of the kind that justifies moral responsibility) in a moment because we never had it in any prior moment. The arguments presented above for real-time control and effortful deliberation undercut this view. If our decision-making is not in fact deterministic and not simply the inevitable consequence of neural or cognitive happenings within us, but something *that we do*, as holistic selves, then RS's primary argument crumbles.

**Shaping our own character**

Moreover, our character (and the configuration of our brain) is not just passively affected by the outcomes of our freely chosen actions. There is ample evidence that we can also act to consciously and deliberatively shape our own character and develop reflective capacities for *moral agency* (e.g., Bandura, 2001; Banicki, 2017; Narvaez, 2019; Narvaez and Lapsley, 2009; Nucci, 2019; Pasupathi and Wainryb, 2010). We learn self-control. We learn prosocial behavior. These are inculcated in us when we're children but they're also things we can take charge of ourselves as we mature. We are able to consider the kinds of motivations we think we should have. We can develop our moral practices and policies and meta-policies in a mindful, conscious way, through time – developing these capacities as skills, not just as dispositions forced on us by accumulated circumstance. Thus, many aspects of our character are indeed *up to us*. It is precisely these kinds of "self-forming actions" that some philosophers (ancient and modern) take to be the essence of our free will, not just in any given instant, but *through time* (Cicero, 1913 translation; Frankfurt, 1972; Kane, 2011; Lemos, 2015; Sedley, 1983).

Thus, neither the claim that we have no real control in the moment, nor the claim that we have no active involvement in the development of our own character are well supported. This is not to say that we ever act free from any prior cause whatsoever – that is an incoherent notion. It is just that we ourselves are causally involved in events, *as agents*, and not just as the site of complicated happenings. This provides a foundation for a naturalistic, non-magical, non-dualist conception of free will, as referring to the evolved suite of capacities that allow us to guide our own behavior (Mitchell, 2023; Steward, 2012; Tse, 2013). These capacities clearly vary in important ways between individuals (Fletcher and Carruthers, 2012; Friedman and Miyake, 2017; Mitchell, 2018) – indeed, this is

part of the evidence for their existence. And it is clearly important to understand the nature and limitations of these capacities for any discussion of moral responsibility and moral desert more generally. But there is no good reason to begin such a discussion with an absolutist (and unsupported) metaphysical claim that shuts down the very debate RS seems to want to have.

**Note:** some of this text is reproduced or adapted from these previously published blogposts:

http://www.wiringthebrain.com/2024/10/the-justice-algorithm.html
http://www.wiringthebrain.com/2024/01/undetermined-response-to-robert_22.html

**Acknowledgments**

Thanks to Henry Potter and Magdalena Paluchowska for very helpful comments on the manuscript.

**References**

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in cognitive sciences*, *21*(8), 607-617.

Artinger, F. M., Gigerenzer, G., & Jacobs, P. (2022). Satisficing: Integrating two traditions. *Journal of Economic Literature*, *60*(2), 598-635.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual review of psychology*, *52*(1), 1-26.

Banicki, K. (2017). The character–personality distinction: An historical, conceptual, and functional investigation. *Theory & Psychology*, *27*(1), 50-68.

Bossaerts, P., Yadav, N., & Murawski, C. (2019). Uncertainty and computational complexity. *Philosophical Transactions of the Royal Society B*, *374*(1766), 20180138.

Cicero. (1913). *De Officiis* (Walter Miller, trans.). Cambridge, MA: Harvard University Press.

Cisek, P. (2019). Resynthesizing behavior through phylogenetic refinement. *Attention, Perception, & Psychophysics*, *81*, 2265-2287.

Cisek, P., & Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual review of neuroscience*, *33*(1), 269-298.

Del Santo, F. (2021). Indeterminism, Causality and Information: Has Physics Ever Been Deterministic? In A. Aguirre, Z. Merali, & D. Sloan (Eds.), *Undecidability, Uncomputability, and Unpredictability* (pp. 63-79). Springer International Publishing.

Del Santo, F., & Gisin, N. (2019). Physics without determinism: Alternative interpretations of classical physics. *Physical Review A*, *100*(6), 062107.

Ellis, G.F.R. (2016). How can physics underlie the mind. *Top-Down Causation in the Human Context.* Berlin and Heidelberg: Springer-Verlag.

Faisal, A. A., White, J. A., & Laughlin, S. B. (2005). Ion-channel noise places limits on the miniaturization of the brain's wiring. *Current Biology*, *15*(12), 1143-1149.

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: computation, biology and function. *Philosophical transactions of the royal society B: Biological sciences*, *367*(1594), 1280-1286.

Fletcher, L., & Carruthers, P. (2012). Metacognition and reasoning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1366-1378.

Frankfurt, H. (1972). Freedom of the will and the concept of a Person. *Journal of Philosophy*, 68(1), 5–20.

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*, *86*, 186-204.

Gigerenzer, G. & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review Psychology*, Vol. 62, pp. 451-482

Glimcher, P. W. (2004). *Decisions, uncertainty, and the brain: The science of neuroeconomics*. MIT press.

Glimcher, P. W. (2005). Indeterminacy in brain and behavior. *Annu. Rev. Psychol.*, *56*, 25-56.

Jaeger, J., Riedl, A., Djedovic, A., Vervaeke, J., & Walsh, D. (2024). Naturalizing relevance realization: why agency and cognition are fundamentally not computational. *Frontiers in psychology*, *15*, 1362658.

Kahneman, D. (2011). Thinking, fast and slow. *Farrar, Straus and Giroux*.

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

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). New York: Oxford University Press.

Lemos, J. (2015). Self-forming Acts and the Grounds of Responsibility. *Philosophia*, *43*, 135-146.

Lemos, J. (2021). The Indeterministic Weightings Model of Libertarian Free Will. Journal of Philosophical Theological Research, (special issue on Free Will), 23(89), 137-156.

Mitchell, K. J. (2023). Free agents: how evolution gave us free will. Princeton University Press.

Mitchell, K. J. (2018). Innate: How the Wiring of Our Brains Shapes Who We Are. Princeton University Press.

Mitchell, K. J., & Potter, H. D. (2024). Beyond mechanism–extending our concepts of causation in neuroscience.

Narvaez, D., & Lapsley, D. K. (2009). Moral identity, moral functioning, and the development of moral character. *Psychology of learning and motivation*, *50*, 237-274.

Narvaez, D. (2019). Moral development and moral values: Evolutionary and neurobiological influences. In D. P. McAdams, R. L. Shiner, & J. L. Tackett (Eds.), *Handbook of personality development* (pp. 345–363). The Guilford Press.

Nicholson, D. J. (2013). Organisms≠ machines. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *44*(4), 669-678.

Nucci, L. (2019). Character: A developmental system. *Child Development Perspectives*, *13*(2), 73-78.

Pasupathi, M., & Wainryb, C. (2010). Developing moral agency through narrative. *Human Development*, *53*(2), 55-80.

Pereboom, D. (2014). The disappearing agent objection to event-causal libertarianism. *Philosophical Studies*, *169*, 59-69.

Potter, H. D., & Mitchell, K. J. (2022). Naturalising agent causation. *Entropy*, *24*(4), 472.

Potter, H. D., Ellis, G.F.R., & Mitchell, K. J. (2025). Reframing the free will debate, Part 1: The universe is not deterministic.

Redish, A. D. (2013). *The mind within the brain: How we make decisions and how those decisions go wrong*. Oxford University Press.

Rich, P., Blokpoel, M., de Haan, R., & van Rooij, I. (2020). How intractability spans the cognitive and evolutionary levels of explanation. *Topics in cognitive science*, *12*(4), 1382-1402.

Rusakov, D. A., Savtchenko, L. P., & Latham, P. E. (2020). Noisy synaptic conductance: bug or a feature?. *Trends in Neurosciences*, *43*(6), 363-372.

Sanborn, A. N., Zhu, J. Q., Spicer, J., León-Villagrá, P., Castillo, L., Falbén, J. K., ... & Chater, N. (2024). Noise in cognition: bug or feature?. *Perspectives on Psychological Science*.

Sapolsky, R. M. (2023). *Determined: Life without free will*. Random House.

Schopenhauer, A. (1960). *Essay on the Freedom of the Will*. New York: Dover

Sedley, D. (1983). Epicurus' refutation of determinism. *SUZETESIS*, 11–51.

Shadlen, M. N., & Kiani, R. (2013). Decision making as a window on cognition. *Neuron*, *80*(3), 791-806.

Simon, Herbert A. (1956). "Rational Choice and the Structure of the Environment". *Psychological Review*, 63 (2): 129–138.

Simon, H. A. (1990). Bounded rationality. *Utility and probability*, 15-18.

Smolin, L., & Verde, C. (2021). The quantum mechanics of the present. *arXiv preprint arXiv:2104.09945*.

Sterling, P., & Laughlin, S. (2015). *Principles of neural design*. MIT press.

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

Tse, P. U. (2013). *The Neural Basis of Free Will: Criterial Causation*. The MIT Press.

van Strien, M. (2021). Was physics ever deterministic? The historical basis of determinism and the image of classical physics. *The European Physical Journal H*, *46*(1), 8.

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.