

Varieties of free will and agency

Wednesday 1 July 2026

**A symposium of the 2026 convention of the UK Society for the Study of Artificial
Intelligence and the Simulation of Behaviour (AISB)**

AISB 2026, University of Sussex, Falmer, Brighton, UK

Varieties of free will and agency

What does it mean to say that a certain class of agents have free will? This year's symposium was a follow-up to last year's successful AISB symposium.

Free will is broadly considered to be an agential capacity, tied to attribution of responsibility both moral and legal. Is free will an illusion? If so, is it one over which – lacking free will – human beings have no control? If free will is real, who has it and what does it mean to have it? Is free will limited to human beings, or do certain other animal species have it? Is it conceivable that a present or future artefact could have it?

What are the properties that mark a given agent as having free will or not? Must a free-willed agent be by some appropriate description intelligent? Must it be conscious? Must it be reflectively self-aware? Some would ascribe all these properties to present-day chatbots. Do they have free will? If not, what are they lacking?

Is free will dependent on “the ability to do otherwise”, as leeway compatibilists and many if not most libertarians would have it? Is it rather dependent on proximal locus of control, as Daniel Dennett has suggested? Does free will merely require the agent being the source of its own agency — the agent's actions are in some substantive sense its own — as the source compatibilists claim? Is it the agent's history that matters: how the agent came to acquire its properties of agency? Do purported AI agents have these properties, or are they “agents” in a different, perhaps strictly metaphorical sense?

How much, if at all, does it matter if an agent “really” has free will or not? What is the difference between an agent claiming free will and actually having it?

What would it mean for a non-human species or an artefact to be successfully identified as having free will? How would that change its relation to us, and how might that change our understanding of what it means to be human?

The contributions to this symposium addressed one or more of these questions. We hope you enjoy reading the papers and abstracts!

Joel Parthemore
Robyn Waller
Nadine Elzein

Papers and abstracts

Thinking, doing, and doing by thinking: NeuroAI and agency (Robyn Waller).....	4
Why libertarians should adopt a narrative view of personal identity to mitigate the luck objection (Alessandro Fiorello).....	5
Fundamentalist libertarianism and emergentist libertarianism (Daniele Conti).....	13
Chuck the cookbook: Why instrumentalists should abandon formally specified desert (Kristoffer Moody).....	14
Metaphysical freedom, political freedom and the empirical-authentic self distinction (Tam Winter).....	16
What is a prediction, anyway? A paradox of everyday agency and scientific prescience (Twan Kieboom).....	19
On moral responsibility, free will and intentional agency in AI systems (Paulius Rimkevičius)	26
Graded minds, graded agency: Why panpsychism cannot exclude AI from the domain of free will (Mehmet Mirioğlu).....	28
Three wills worth wanting: Source freedom and second wave AI (Daniel Haas).....	38
What's the point? Indeterminist compatibilism and the implications for agency and AI (Simon Bowes).....	42

Thinking, Doing, and Doing by Thinking: NeuroAI and Agency

Robyn Repko Waller¹

Abstract.

This talk introduces and explored an overlooked metaphysical tension in the philosophy of neurotechnology: It seems, *prima facie*, intuitive to endorse that we can do by thinking and that neurotechnology, such as BCI, allows us to do by thinking in novel, transformative ways. This is agency optimism about BCI technology. However, attention to agency-undermining interpretations of experimental decoding of intentions in the neuroscience laboratory (e.g., interpretations of Libet-inspired paradigms) suggests a pessimism regarding the exercise or restoration of agency with BCI technology.

If the parallels between the methodology and data of the Libet paradigms and BCI technology hold, then we ought to conclude the same thing metaphysically about agency in both cases. Do decoding algorithms and related neurotechnology hold promise to harness and expand our agency or, on the contrary, merely cast an illusion of control? The talk outlines the best metaphysical arguments for agency optimism about BCI and pessimism about BCI agency and gestures at how further argumentation might establish a kind of optimism, metaphysically, that BCI users are exercising agency.

Why Libertarians Should Adopt A Narrative View of Personal Identity to Mitigate The Luck Objection

Alessandro Fiorello¹

Abstract. Robert Kane's libertarian theory of free will is the most widely influential in the philosophical literature in the past fifty years. The most perennial objection to Kane's account has been the luck objection. Alfred Mele's articulation of the luck objection is the most influential as it captures the spirit of Kane's critics. According to this objection, if self-forming actions (SFAs) are undetermined, then nothing about the agent contrastively explains why one outcome rather than another occurs, rendering the choice a matter of luck and the agent as not the ultimate source of her SFA. I argue that we do not need a contrastive explanation for an agent to be an ultimate source of her SFA. Ultimate sourcehood is preserved by grounding it in a diachronic conception of narrative authorship and personal identity. On this view, agents are ultimate sources of their actions not by settling outcomes through decisive reasons, but by authoring practical identities over time through self-forming commitments made under conditions of indeterminacy. Narrative identity explains how SFAs can be attributed to agents even when their outcomes are not contrastively explicable. I conclude that once ultimacy is properly understood as diachronic authorship rather than momentary control, the luck objection fails to undermine Kane-style libertarian free will.

Keywords: Libertarian Free Will; Moral Responsibility; Narrative Identity; Moral Luck

1 INTRODUCTION

Libertarian theories of free will and moral responsibility are the conjunction of the following theses:

- (1) Free will is incompatible with determinism
- (2) Either determinism is true and we lack free will or we have free will and determinism is false
- (3) We have free will
- (4) Determinism is false[1,2,3]

Libertarianism is committed to the incompatibility of free will with determinism and affirmation of the positive portion of the disjunction in (2) which entails (3) which from (1) entails (4). Libertarianism holds (1) because it takes as a necessary condition for free will the ability to do otherwise. [4] This condition is inconsistent with the truth of determinism. It is not possible for determinism to obtain and for an agent to have done otherwise holding fixed the past and the laws of nature. Libertarian theories of free will and moral responsibility face a perennial problem called the luck objection. Nadine Elzein notes the problem comes

in more than just one form. I am going to focus on one specific form of the problem from luck.² This branch of the problem of luck asks us to consider explanations for undetermined actions and why they happened as they did as oppose to how they could have gone. The problem demands that actions must be explainable in a contrastive fashion in order to attribute control to agents over undetermined actions. What is a contrastive explanation? A general and noncontroversial definition of contrastive explanations is that they are explanations where the explanandum (the fact to be explained) is why x rather than y. X is the fact and y is the foil that is contrasted with x. An example from the literature constructed by Fred Dretske helps to elaborate this definition.[6] The General turns the key in the ignition of his car and the car explodes. A contrastive explanation asks why turning the key caused the explosion rather than starting the car. We can ask the same question for undetermined actions. Why did Jane choose to go to law school rather than go to graduate school? The worry as far as indeterminism goes is that we are unable to explain why Jane chooses one option rather than another so it is just an inexplicable happening. The most influential form of this species of argument is put forward by Al Mele.[7] Robert Kane is the most influential libertarian in the past five decades and has tried to construct a persuasive reply to Mele but has not succeeded.[10,11,12,13,14,15,16,17,34,35,18,19,9,21,22] Kane's theory argues that agents must be ultimately responsible for shaping their wills. In order to satisfy the condition of ultimate responsibility an agent must make what Kane calls self-forming actions or SFAs. SFAs occur when an agent is motivationally torn between alternatives that their will is not set or torn between making.

In this paper, I attempt to offer a reply that mitigates the force of this objection to libertarianism. I am going to construct a libertarian account that builds and expands upon Robert Kane's libertarian model of free will.[16] The key pillar in Kane's view is his idea of self-forming actions or SFAs which entail agents making undetermined choices that shape their wills. I am going to expand on Kane's view through the idea that agents are metaphorical authors of their metaphorical life stories. I argue that self-formation occurs in a broader narrative framework and when we see SFAs as part of the stories of our lives, they are not just a matter of luck. When agents make an SFA, they are authoring their life story and they endorse whatever they decide because the outcome of SFAs are the continuation of their life story, no matter the outcome. A narrative view of personal identity allows libertarians like Kane to offer a non-contrastive explanation of undetermined actions like SFAs that feature aspects of the agent.

¹University of Ottawa

²Elzein discusses alternative instances of the problem of luck such as “the disappearing agent” objection as well as claims that indeterminism has to be random and uncaused [5].

Narrative authorship shows that agents are the ultimate source of their SFAs even though their choice is not contrastively explainable. When an agent makes an SFA, they are also making what Kane calls a value experiment.[16] If we cast the explanation of SFAs as value experiments that take place in a wider teleological or narrative structure, we can see that SFAs are not just a matter of luck. Therefore, luck does not vitiate ultimate responsibility.

My argument is constructed as follows: section two outlines Kane's model and his reply to the luck objection. Kane's response to the luck objection relies on what he calls plural voluntary control. It makes important progress on the issue but it has not been seen as completely resolving the issue because it still does not provide a contrastive explanation. In section three I provide Mele's luck objection. Mele argues that Kane is still not able to overcome the luck objection and provides examples that demonstrate Kane has not answered the objection. In section four I argue that a narrative view of personal identity through the concept of authorship can provide a libertarian theory like Kane's a way to mitigate the force of the luck objection. I show that even if we are not able to contrastively explain SFAs, we can provide a narrative explanation of SFAs that stops luck from vitiating moral responsibility. I conclude in section five that I have shown libertarians are able to undercut the force of the luck objection through narrative authorship and adopting a narrative view of personal identity.

2 KANE'S MODEL OF SELF-FORMATION

Kane's theory is best captured through his example of a businesswoman on her way to an important meeting.[16,23] While on her way to her meeting she sees an assault taking place in an alley and an inner struggle occurs between her conscience and her career ambitions.[23] She can intervene and call for help, or she can keep ignore the person in need of help and continue on to her meeting. She has reasons to which she is responsive such that both options are motivating and so entail this inner conflict. The resolution of such conflict involves what Kane calls, "will-setting," whereby agents incline their wills in their moments of inner conflict by deciding.[24] We set our wills through forming them by performing SFAs. Thus, the businesswoman's will-setting action is also an SFA. By setting our wills via SFAs we satisfy the conditions for what Kane calls ultimate responsibility or UR.[16] Satisfying the conditions of UR entails that agents are responsible for not only what they will to act but for the will itself from which they act. This does not entail fully self-forming one's will but rather, that in some significant sense, our wills are 'up to us' in that they are not completely necessitated by prior factors beyond our control. She satisfies the conditions for UR by having a will shaped through her SFA.

Kane's account argues for a unique conception of indeterminism. According to Kane, indeterminism acts a hindrance or obstacle in SFAs because agents are engaged in an internal struggle between conflicting values internal to an agent's

will. When an agent makes an SFA, they are torn between competing motivations and they must make mental efforts to overcome the temptation to choose the alternative possibility. Indeterminism functions as a hindrance in the sense that agents face live options that are genuinely possible to choose between because their action is undetermined.[13] Since the businesswoman is able to do otherwise in an SFA she must overcome the temptation of going to her meeting to help the assault victim, and she must overcome the temptation of helping the assault victim to get to her meeting. She is internally conflicted because her will is divided as a result of values internal to her will pulling her in incompatible directions. It is important to note that indeterminism is an ingredient in larger goal-directed processes of the agent in Kane's model.[13] It is significant to point this out because when an agent makes an undetermined choice like an SFA, their decision is in spite of indeterminism and not because of it. Indeterminism engenders a metaphysically open future because it makes the internal struggle between competing values a struggle between genuine alternative possibilities.³

This brings us to Kane's response to the luck objection. Kane claims that agents like the businesswoman have what he calls plural voluntary control (PVC), which he uses to combat the luck objection.[16] When an agent makes an SFA they exercise PVC. PVC is the kind of control exercised over an action by an agent when they are able to choose intentionally, rationally, and on purpose and can do otherwise. The businesswoman exercises PVC because she can do otherwise intentionally, rationally, and on purpose. PVC provides an explanation for SFAs that Kane thinks is able to answer the luck objection. The businesswoman acts for reasons that Kane calls satisficing reasons. These are reasons which are not conclusive or decisive but "good enough" to justify a choice. PVC gives the businesswoman control over her SFA because she has satisficing reasons for both alternatives and she intends to do both for those reasons. The businesswoman's choice will not therefore be just a matter of luck. Despite the indeterminism, there is a causal link between the businesswoman's decision and her reasons. When we cast an explanation of SFAs in terms of a larger indeterministic causal framework he thinks the luck objection is answered. Whatever the businesswoman decides there will be a causal link between her decision and the outcome. Kane believes it is just false that SFAs are just a matter of luck because the luck objection fails to properly understand how indeterminism works alongside PVC. Despite Kane's reply here to the luck objection it is still a problem for his model.

Kane argues that when we make SFAs, we are also initiating what he calls value experiments. He says that every SFA is a value experiment, "whose justification lies in the future and is not fully explained by the past. It says, in effect, "Let's try this. It is not required by my past, but it is consistent with my past and is one branching pathway my life could now meaningfully take. I'm willing to take responsibility for it one way or the other. It may make me a (morally) better or worse person, a (prudentially) happier or sadder person, but I will be aware that it was my doing and my own self - making either way---guided by my past, but

³Kane believes that indeterminism is a necessary condition for free will because it makes the internal struggle in our wills create a genuine open future leaving us to freely choose what future to bring into existence. This internal struggle is compared to the struggle birds face with the

wind. In order for a bird to fly they require the resistance of the wind. In order for human beings to be self-formed, we need to struggle and achieve the values we endeavour to create and sustain.

not determined by it." [16] Kane argues that value experiments take place within a wider teleological or narrative framework. As a means of exemplifying the idea consider Bernard Williams' fictionalized example of Paul Gauguin. Gauguin is a married man with children who works as a banker. He also aspires to be a great painter. One day, an opportunity presents itself for Gauguin to leave France for Tahiti where Gauguin can pursue his artistic ideals. Gauguin becomes conflicted over whether he should leave France and his family behind or if he should abandon his goal of becoming a great painter. Gauguin's past provides reasons which are not sufficient but rather they are what Kane calls *satisficing* reasons, which he defines as "reasons that are good enough to justify a choice or action even though they are not sufficient to render any possible alternative choice or action that might have been made in the circumstances unreasonable or irrational." [25] Gauguin's past does not fully explain his present or point towards one single option as the only reasonable one. So, Gauguin conducts a value experiment where he chooses for satisficing reasons. The way that Gauguin's life has unfolded up to that point when taken as whole leaves open his future in regard to his reasons not conclusively supporting one course of action rather than another. When Gauguin deliberates over his situation, he takes a look at his life as a whole; his present beliefs about his future possibilities are shaped by his past. The choice he makes will be self-forming as the way he chooses to continue his life story will impact his will and through making such self-forming choices, he is the author of his own life story. Kane does not offer an account of the interconnection between a narrative view of personal identity and his view of free will but merely alludes to it. [25] I think Kane is correct to identify a connection between narrative and self-formation but I am going to develop a more extensive view of narrative self-formation that greatly expands and builds on Kane's libertarianism in section five.⁴

3 MELE AND LUCK

In his essay, "Kane, Luck, and the Significance of Free Will," Alfred Mele argues that Kane has not answered the luck objection, and he appears to speak for many of Kane's critics. [7] Mele motivates this claim with a discussion of indeterminism. He writes,

Libertarians rightly regard determinism as a bar to agents' having more than one causally open future (i.e., more than one future that is consistent with their past and the laws of nature), ... Although indeterminism removes this bar, it also raises a worry about luck. If there are causally undetermined or indeterminate aspects of a process that terminate in, for example, a choice-- including, as in Kane's picture, aspects that are present at the very time the choice is made and directly relevant to the process's outcome--then, to the extent that the agent is not in control of these aspects, luck enters the picture in a way that seemingly threatens both moral responsibility and any desirable species of free agency. [7]

Mele believes Kane's account does not provide enough of the control required to be responsible for making an SFA. He uses a thought experiment about an agent named Ann to demonstrate his

point. Ann has made a promise to call her colleague Beth at eight in the morning to prepare her for an important meeting. Beth has recently broken a promise to Ann and Ann is still upset over it. Ann believes she ought to call Beth on time but she also wants revenge. She is conflicted and engages in an SFA where she chooses to get revenge and does not call Beth on time and waits several minutes. Consider an agent in an identical possible world and who is identical to Ann--call her Ann*--who faces the same situation. She too faces the same internal conflict but is able to overcome her desire for revenge and calls Beth on time. [7] Mele challenges the libertarian to explain this difference in outcome. Mele wants an explanation that cites the, "powers, capacities, states of mind, moral character, and the like that explains the difference in outcome" Without an explanation that uses something along these lines "then the difference really is just a matter of luck." [7]

Mele will not be satisfied by Kane's appeal to PVC because Kane rejects contrastive explanations for SFAs and uses PVC to try to offer a non-contrastive explanation. PVC will allow us to say Ann and Ann* *make choices that are rational and intelligible and so it is not just a random luck of the draw why do what they choose to do. However, given that Ann and Ann* are sufficiently similar and they are able to do otherwise, the question that created the problem still remains. Why do Ann and Ann* choose differently from each other? If we jettison contrastive explanations, then we seem unable to provide an answer for the difference in outcome that is something other than luck. If we are not able to pick out some feature of Ann and Ann* that contrastively explains the difference in outcome, then the only explanation available appears to be one got morally lucky while the other got morally unlucky. This seems like a highly pertinent concern because both agents are supposed to be able to have done otherwise than they actually did. If Ann and Ann* have actual alternative possibilities, then why do they choose differently from one another. Mele believes we need a contrastive explanation that highlights some feature of Ann and Ann* to show it is not just a matter of luck that they choose as they do.*

I disagree that we need a contrastive explanation in order to demonstrate that Ann and Ann* are ultimately responsible for their SFAs. I think it is possible to provide a noncontrastive explanation that makes clear SFAs are not just a matter of luck. Through showing that SFAs are made in the context a wider narrative of the agent I will show that SFAs are not just a matter of luck.

4 CONTRASTIVE EXPLANATIONS, LUCK, AND ULTIMATE RESPONSIBILITY

The luck objection seems compelling, but I will argue that it assumes that luck necessarily vitiates responsibility. If an agent's action is produced by luck, then they cannot be held responsible for performing it. Call this luck incompatibilism. Luck incompatibilism seems to be the thesis that an agent's free will and moral responsibility are incompatible with the occurrence of an agent's action involving luck. When put under scrutiny luck incompatibilism falls short. The reason is that there is a difference between something happening solely due to luck, and luck being

⁴I should also mention that for lack of space I won't be able to discuss Kane's own attempt to solve the challenge

to his view raised by Mele. For those interested, see [16], ch. 6–8.

part of an overall process. Luck incompatibilism relies on the former being the case, but I argue that narratively structured self-formation involves the latter. The previous section shows that Mele's luck objection to Kane is that without a contrastive explanation SFAs are just a matter of luck. Therefore, agents are not ultimately responsible for their SFAs because they just get lucky. Luck vitiates ultimate responsibility and is incompatible with either free will or ultimate responsibility. It seems fair to say then that luck incompatibilism relies on the following requirement:

(Contrastive Explanation Requirement or CR) If an agent freely chooses A rather than B at time t , then there must be some agential feature present at t that explains why A occurred rather than B.

CR brings forth what luck incompatibilism assumes to be true and it is quite clearly at work in Mele's articulation of the luck objection. Iff an explanation that contrastively explains an agent's choice in terms of some feature of the agent is available is that choice not just a matter of luck. This is not a neutral, trivial or esoteric principle or criterion. CR plainly states that for an agent to be responsible for their action, there must be something about them that is sufficient to decide why A occurred rather than B. In other words, there has to be a sufficient reason that determines why one event occurred rather than another. Since SFAs are undetermined actions they cannot have such sufficient reasons. Luck incompatibilism through CR would lead one to conclude that since SFAs cannot satisfy CR they must be just a matter of luck.

I am going to show that self-formation incorporates an element of luck in a larger goal directed process that is part of a larger narrative, and that agents do not make their SFAs only because of luck. I am going to begin with showing that an agent is still responsible for their choices even when luck is involved in their decisions. I will then apply this lesson to SFAs.

Consider the following two cases about two individuals named Alfred and Brenda:

Alfred uses a coin toss to decide between two options. The outcome of the toss is either heads or tails---this is not up to Alfred. But he endorses this method of choosing, including the outcomes that result. And he initiates the operation of this way of choosing---i.e., he chooses when to use this method, and he tosses the coin. Brenda lives in Ottawa. She decides to live one whole year there wearing nothing but short-sleeved shirts and shorts (and sneakers on her feet). Part of her reason for doing this is to experience the effects of harsh winter weather on such a slightly dressed person. She doesn't know how cold the winter will be exactly, and she doesn't know if she'll have to go out on the worst of days. What the weather will be is not up to Brenda, nor are the effects on bare skin. But she endorses the exposure that she will undergo, and to a certain extent her engagement with the weather is up to her, since she can avoid going outdoors sometimes.

While Alfred and Brenda are engaged in activities which involve luck the outcomes of their decisions---for example Alfred's choice of hamburger for dinner using the coin toss, Brenda's discomfort on some days, but avoidance of frostbite on others, even when she goes outdoors---is attributable in some sense to Alfred and Brenda, maybe even wholly. The fact that we

can attribute the outcome of these cases to the agents involved suggests, I think, that an agent can be ultimately responsible for their action even when luck is involved in the production of the action. Alfred and Brenda do not choose as they do because of luck but rather their choices involve luck and there is distinct difference between the two. The difference stems from the relationship one has to luck.

The nature of this relationship that allows for Alfred's action to be attributable to him is one of endorsement. Alfred endorses both using this tiebreaking method to choose and when he uses this method to make decisions. Similarly, Brenda endorses her exposure to cold weather and obviously has no control over the weather. Yet her willingness and endorsement of exposure to cold weather makes this activity attributable to her. Alfred's circumstances involve choosing a tie-breaker when his reasons are not conclusive on their own. His endorsement of the outcome of the coin toss suffices for the action to be attributable to him as the source of it. He chooses this method, and while the coin toss itself is not within his control, his endorsement of the outcome and the coin toss seem to make him the source of the outcome.

Alfred and Brenda's endorsement places them in the right relationship to the lucky processes they are in relationship to because of their willingness to use these processes. Alfred and Brenda both act for reasons that are voluntarily, intentional and not produced by coercion. While the outcome of their activities is subject to luck, they willingly do these things and the lucky process involved produces things they each endorse. The reasons involved in the case are separate from the endorsement, but they are relevant to the description of the cases. They are important for demonstrating that Alfred and Brenda are acting freely in the sense that they are acting intentionally and on purpose; it is significant for establishing that their actions are not just an accident or purely random. The role of endorsement in their respective cases is to show how the action they perform is still attributable to them in the sense of ultimacy. They endorse the outcome of the lucky process they choose and endorse utilizing lucky processes. The endorsement of the agent is the bridge between agency and luck that allows the compatibility of ultimacy and luck.

Alfred and Brenda's endorsement demonstrates their willingness to undergo activities---the coin toss in Alfred's case and Brenda's exposure to cold weather while being underdressed---which involve luck. They are fully aware of the luck inherent in their actions and willingly pursue their goals with their respective lucky action processes. In this relationship between agents and luck, agents can see that the outcome of their actions are uncertain and open to being influenced by luck. Nonetheless, the outcome is something they want to do and so they endorse the outcome even though luck is involved in its production. It seems to me the kind of luck in cases like Brenda is of the nature that agents can be fully aware of its presence and fully willing to accept its presence and role. Coin tosses and the weather are things whose outcomes mature agents are aware are beyond their control but still willingly engage with or use to settle their minds. Brenda knows full well what Ottawa is typically like in the winter and she does not control the weather but still chooses to go out in summer clothes. Alfred does not control the outcome of the coin toss but endorses whatever outcome it 'decides.' He need not use a coin toss, he could have chosen some other means he might have control over. He intentionally uses a coin toss to decide for him. That these agents willingly, voluntarily and intentionally engage in

these lucky processes makes it appropriate to attribute what they do as ultimately up to them and ultimate sources of their behavior.

Imagine that instead of Alfred and Brenda endorsing the use of lucky processes, they merely acknowledge that luck is involved. Suppose Alfred uses the coin toss but he does not endorse using this as a method but indifferent. Alfred is simply lazy and uses a coin toss for convenience. Alfred is aware that by using a coin toss he opens himself up to luck. Through his awareness that using a coin toss subjects him to luck and choosing to go through with it, this makes Alfred in the right relationship to luck.

An agent can be in the right relationship to luck here without an awareness or concern about luck. All that would be needed is a view of oneself and one's actions as subject to luck to be incorporated (perhaps in specific ways) into one's sub-personal self-image-producing and action-producing capacities. In this case, we have the possibility of all normal people acquiring such a self-image, and hence the right relationship to luck in general, through typical maturation. The (by hypothesis uncommon) absence of such a sub-personal self-image might be detectable by things like strange sorts and/or degrees of surprise about worldly processes that affect one's actions in unusual ways, or lack of flexibility when the world at large doesn't meet one's expectations. Such responses (behavioral, attitudinal) would mark a dissociation between one's view of oneself and one's understanding of the wider world and the contingency found therein.

The advantage of this view is the possibility of all normal people developing the right relationship to luck. This seems like a virtue of an account that seeks to make ultimate responsibility compatible with luck. This also opens the possibility that any normal people that don't develop this sort of relationship to luck could end up in the relationships described in the first two cases. Indeed, if one saw one as subject to luck but disavowed it in the maturation of oneself, this would put an agent in the second sort of relationship. If one came to be aware and endorsed luck, an agent would be in the first sort of relationship.

Suppose Amara is a maturing agent who is making a choice between rationally satisficing reasons. This introduces luck into her deliberative and action-producing processes. For Amara to stand in the right sort of relationship to luck she need not be aware, or care about luck. What is needed is a view of herself as being open to not luck per se as this would bring luck to a level of awareness not required. Instead, Amara needs to come to incorporate luck into her sub-personal self-image through the development of a self that sees herself as open to failure and an uncertainty about what to do and a willingness to choose, nonetheless. What I am trying to capture here is that Amara sees her choice as being open to not meeting her expectations and that she doesn't know which choice is the right one to make. The choice she ends up making might be subject to luck, but from the first-personal perspective she doesn't see herself as making a lucky choice. She sees herself as making a decision she is not certain is the right one but which she willingly and intentionally chooses. At the sub-personal level her action-production might involve luck, but this does not vitiate the attribution of her choice

to Amara. Amara's choice still comes from her in that she is choosing for reasons that motivate her and this would be true no matter what Amara does. That luck is involved seems besides the point to Amara developing into a mature agent which, in the view of my theory, is an ultimately responsible self-author. Amara, at the level of her awareness, is making choices that are importantly connected to her will: her values, desires and reasons are all causal contributors to the production of her choice. In spite of the luck at work here, Amara is still an ultimate source in virtue of the contributions of herself in the form of these agential features make. The outcome of this process would have been subject to luck but as long as Amara is involved in this manner, she has the necessary relationship to luck to say she is ultimately responsible.

5 LUCK AND NARRATIVE SELF-FORMATION

We have seen that an agent can still be responsible for her decision even when luck is involved in producing the outcome. In this section I am going to expand on this point and provide a narrative account of self-formation. In constructing my view I will incorporate Kane's idea of a value experiment. Here is a helpful sketch of the argument for this section: SFAs occur within a larger narrative unity and when we understand agents as authors engaging in a value experiment about how to continue their life story it becomes clear that SFAs are not just a matter of luck. Even though we suppose that luck is involved in the outcome of an SFA, the narrative structure and unity of an agent provide the right relationship to SFAs to preserve an agent's ultimate responsibility. This is because agents author their narratives through engaging in SFAs which are also value experiments. Since the outcome will be an intelligible continuation of an agent's life story they endorse the outcome no matter what it turns out to be.

A narrative view of personal identity is best captured in terms of a narrative structure.⁵⁶ Human beings have a past, present and future and they are all importantly related to each other in the story of one's life. Having a narrative identity is properly understood to be that one's present sensibly emerges from one's past and what one expects to happen in the future is tied to one's present. The past, present and future of a life are able to be tied together into a single unified narrative structure. Consider Pam who is torn between going to the local community college to be close to friends and family or moving away to a university in a large city which is far away. She has present alternatives about her future that derive their meaning and motivation from her past. Her conflicting desires over whether to stay close to home or to venture to a far away place are embedded in her narrative structure as they are part of her character. Facing a branching pathway into the future Pam is divided between how to extend her present into the future because of how her life trajectory has unfolded at this point. It never occurs to Pam to abandon her pursuit of post-secondary education and become a Buddhist. The explanations of what Pam does can use her character to explain why they do not pursue those alternatives. Her character fail to give conclusive

⁵There are many accounts of narrative in the philosophical literature and I do not have the space or scope to go over them here. See [26–29].

⁶I am aware of critics like Galen Strawson who rejects narrative views of personal identity. While my paper

heavily relies on this approach I do not have the space in this paper to engage with critics like Strawson and provide a rebuttal. So, I ask that this view of personal identity be assumed to be true for the sake of argument. For Strawson's critique see [30].

reasons but instead supply satisficing reasons. Pam's reasons for choosing the local college or choosing the university far away incline her without necessitating her choice because neither is sufficient for choice. That is to say, whatever Pam's decision, her character and past intelligibly lead to either decision being reasonable, but neither is definitely better than the other.

The ambiguity over how Pam could intelligibly and coherently continue her life story as a whole shows how sometimes examining our lives holistically shows how they have a narrative structure. On my view of narrative, persons are not writing a life story towards any specific goal, purpose, or end. Narratives serve persons, not the other way around. The narrative structure of our lives is a tool that can be used to help lead good lives, not necessarily creating a good story. Pam may think of her moment of self-reflection as writing a story but that is not required. Her past and future are pulled together in the present and are linked together in such a way as to constitute a single, unified narrative. Suppose Pam is making an SFA. Since Pam is making a SFA she is also engaged in a value experiment. Here is how this matters to free will. The discussion of narrative and narrative identity points to a view of free will in terms of authorship. The exercise of free will involves, among other things, the authoring of our identities and our lives through the initiation of value experiments that are contextualized in the narrative structure of our lives. This broader narrative structure contains a powerful response to mitigating Mele's compelling objection from luck. The challenge posed by Mele's concern about luck is to offer an explanation of an agent's action in terms of the "total prior character, motives and purposes of the agent." [7,9] Specifically, he wants an explanation which allows us to say why one event happened rather than another. This is a contrastive explanation and it must include personal characteristics in order to say that it is the agent and not luck that is responsible for what the agent does. Using the account of narrative given in the last section, it is possible to show that agents engage in self-formation in a wider narrative structure, and this demonstrates that agents are responsible for authoring their life stories and identities. The way that the use of narrative preserves responsibility is we are able to attribute SFAs to agents because they are the ultimate sources of their actions.

Consider Josh, who is making an SFA and has rationally satisficing options between missing an important business meeting and going to his children's music recital. Suppose Josh misses the recital. Now imagine a counterfactual agent, Josh\, *who has the same past and lives in a universe with the same laws of nature as Josh but goes to the recital. If both Josh and Josh\ are able to do otherwise because their actions are undetermined, then what accounts for the difference between Josh missing the recital and Josh* going to the recital?* Given they are similar enough to make the comparison, what is it about their states of mind, characters and motives and the like which would allow us to say made the difference. Moreover, given that they are able to do otherwise why do act as they do and not act otherwise. Without a contrastive explanation the only explanation seems to be that Josh and Josh star do what they do because of luck.

I think that providing a narrative explanation of SFAs shows that agents are still responsible for their SFAs because they are the authors of them. A narrative explanation of SFAs contextualizes agential features in such a way that we can see that the agent and not luck is the source of their decision to A or B. They are the source of their decision because they are the authors of the narrative in which they make their choice. They are authors of

their narratives because agents can take a single unified perspective of their lives and examine it as a whole and endorse the choice to extend their narrative. Kane describes persons as asking themselves who do I want to become when they make an SFA. This introspective question requires a narrative unity to be intelligibly asked because who one wants to become is tied to who one is which is shaped by one's past. The agent's phenomenological perspective of their SFA views themselves as internally divided between competing visions of the kind of person they want to be which is derived from who they are now. Since SFAs have a narrative continuity to them the agent will take responsibility and endorse whatever choice they decide to make. [16]

As either option is able to maintain narrative continuity agents perform a value experiment. Value experiments are characterized by Kane as decisions whose justifications are in the future and are not fully explainable by an agent's past. Kane says that romantic dates, marriages, career choices, economic and political policies are all examples of value experiments. [31] Each of these ordinary cases involves risk and an uncertainty over the success or failure of the outcome. The nature of such experiments is that they are open to verification through testing any expectations against experience. Value experiments can also be plans of action or ways of life and can be found in many parts of our ordinary every day lives. Kane's Kierkegaardian inspiration for value experiments is seen in that every value experiment may lead to despair. [32,33] The inspiration from Mill is seen in that an underlying motivation about value experimentation is that some truths about the good life cannot be learned from the armchair or a priori but only through practical engagement with the world. Since every value experiment might have disappointing results. A person facing a SFA does not have a clear way to sort out which ones will be successful is to live them out. A qualification of this point is that we can actually engage in value experiments in our mind. Kane says that value experiments can and often, "do take place in the head -- imagining scenarios, considering possible outcomes and consequences of actions in the attempt to decide which plans or intentions are worthy of being pursued in reality." [31] The importance of testing them out through living them is that we are fallible. We need something to test our hypotheses against lest we run the risk of spinning fanciful tales in our heads.

Value experimentation within SFAs takes place within a larger narrative structure, the story of one's life. This is true because one's past has shaped one's present. SFAs are events that occur when our previously formed character and motives are in conflict in the present moment. The options before me are partially circumstantial, but they are also related to my character. It is not as simple as either just circumstance or just character. It is circumstantial, as I am in a situation that requires me to make a choice, but it is also related to character, because the options before me are meaningful in virtue of who I am and who I want to become, as well as other things. We make our characters and identities from the situations we find ourselves in. We weave together our various experiences via value experiments initiated in SFAs to create a life that is good for us. SFAs and therefore value experiments are to be explained within a wider narrative context. This is because the conflict distinctive of self-formation arises from how our life stories have unfolded at that moment. The explanation of value experiments is one that conveys the intelligibility of the choice made within the wider narrative context of one's life story.

We now have everything we need to mitigate the force of the luck objection. We can provide a non-contrastive explanation that preserves an agent's responsibility for her SFA by showing that she is the source of her SFA. I showed that an agent is the source of her SFA because she is the author of the decision in virtue of the value experiment she performs. An agent's experimentation with values internal to one's will is contextual to the wider narrative structure or the broader story of their life. Since all SFAs are value experiments, all value experiment will be endorsed by the agent. Their endorsement makes them the authors of the will they are forming. Being the author of the will makes them attributable for their SFAs. As I argued in the last section, being attributable enables it to be the case that luck can part of an overall process while the agent is still responsible for their action. Once we understand SFAs as value experiments taking place in a wider narrative we can see that luck can be part of the process without undermining ultimate responsibility. For all the agential features Mele demands are present in my narrative explanation of self-formation. True enough, it is not a contrastive explanation but we not need to satisfy the CR condition to undermine luck's vitiating of free will and moral responsibility. Agents through their endorsement are the authors of their SFAs and hence are ultimately responsible for their SFAs. Even though we are not able to provide a contrastive explanation we can still provide an explanation that shows SFAs are not just a matter of luck. Agents are the authors of their wills because they are the authors of their SFAs and this makes the outcome of an SFA attributable to them in spite of the presence of luck involved.

6 CONCLUSION

I have demonstrated that through adopting a narrative view of personal identity SFAs are not just a matter of luck. I have provided a noncontrastive narrative explanation that utilizes agential features to show that agents are the authors of SFAs and their wills. If we contextualize SFAs as value experiments that occur in the story of our lives, we can see how luck could be granted to be part a narrative structure without undermining moral responsibility. Agents author their wills by making SFAs and they are the authors of their SFAs because they endorse the outcome of an SFA. An agent endorses the outcome of an SFA because it is continuous with the narrative structure of one's life. Since narrative continuity is able to be maintained by pursuing any possible alternatives agents take the risk to pursue a set of values or way of life. Agents take responsibility for their choice regardless of the outcome. An agent might come to feel happy or sad about their decision, but they endorse it nonetheless as it was a meaningful choice for them to make. As Kane helpfully puts it:

True enough. But I did have adequate reasons for choosing as I did, which I'm willing to endorse and take responsibility for. If they were not sufficient or conclusive reasons, that's because... I was not a fully formed person before I chose (and still am not, for that matter) ... I am in the process of writing an unfinished story and forming an unfinished character who, in my case, is myself.[14]

REFERENCES

- [1] L. W. Ekstrom. *Free Will: A Philosophical Study*. Westview Press, Boulder, CO (2000).
- [2] L. W. Ekstrom. Introduction. In *Agency and Responsibility: Essays on the Metaphysics of Freedom*, ed. L. W. Ekstrom, pp. 2-3. Westview Press, Boulder (2001).
- [3] C. E. Franklin. *A Minimal Libertarianism: Free Will and the Promise of Reduction*. Oxford University Press, New York (2018).
- [4] G. Strawson. *Freedom and Belief*. Oxford University Press, Oxford (1986).
- [5] N. Elzein. The Demand for Contrastive Explanations. *Philosophical Studies*, 176(5):1326 (2019).
- [6] F. I. Dretske. Contrastive Statements. *The Philosophical Review*, 81(4):411-437 (1972).
- [7] A. R. Mele. Kane, Luck, and the Significance of Free Will. *Philosophical Explorations*, 2(2):96-104 (1999).
- [8] A. R. Mele. Libertarianism, Luck, and Control. *Pacific Philosophical Quarterly*, 86(3):381-407 (2005).
- [9] A. R. Mele. *Free Will and Luck*. Oxford University Press, Oxford (2006).
- [10] R. Kane. *A Complex Tapestry of Free Will: A Philosophical Odyssey*. Oxford University Press, New York (2024).
- [11] R. Kane and C. Sartorio. *Do We Have Free Will? A Debate*. Routledge, New York (2022).
- [12] R. Kane. Making Sense of a Free Will That Is Incompatible with Determinism: A Fourth Way Forward. *Journal of Philosophical and Theological Research*, 23(3):5-28 (2021).
- [13] R. Kane. On the Role of Indeterminism in Libertarian Free Will. *Philosophical Explorations*, 19:2-16 (2016).
- [14] R. Kane. Acting of One's Own Free Will: Modern Reflections on An Ancient Philosophical Problem. *Proceedings of the Aristotelian Society*, 114(1, pt.1):35-55 (2014).
- [15] R. Kane. Free Will and the Dialectic of Selfhood. *Ideas y valores*, 58:25-44 (2009).
- [16] R. Kane. *The Significance of Free Will*. Oxford University Press, New York (1996).
- [17] R. Clarke. *Libertarian Accounts of Free Will*. Oxford University Press, Oxford (2003).
- [18] C. Ginet. Can an Indeterministic Cause Leave a Choice Up to the Agent? In *Libertarian Free Will: Contemporary Debates*, ed. D. Palmer, pp. 15-26. Oxford University Press, Oxford (2014).
- [19] M. Griffith. Why Agent-Caused Actions Are Not Lucky. *American Philosophical Quarterly*, 47:43-56 (2010).
- [20] M. McKenna. Ultimacy and Sweet Jane. In *Essays on Free Will and Moral Responsibility*, eds N. Trakakis and D. Cohen. Cambridge Scholars Publishing, Newcastle upon Tyne (2008).
- [21] A. Roskies. Can Neuroscience Resolve Issues About Free Will? In *Moral Psychology*, ed. W. Sinnott-Armstrong, pp. 103-126. MIT Press, Cambridge, MA (2014).
- [22] G. Strawson. The Impossibility of Moral Responsibility. *Philosophical Studies*, 75:5-24 (1994).
- [23] R. Kane. Libertarianism. In *Four Views on Free Will*, ed. E. Sosa, pp. 5-43. Wiley-Blackwell, Malden (2007).
- [24] R. Kane. Freedom, Responsibility, and Will-Setting. *Philosophical Topics*, 24(2):67-90 (1996).

- [25] R. Kane. The Problem of Free Will: A Libertarian Perspective. In *Do We Have Free Will?*, 1st ed., p. 65. Routledge (2022).
- [26] M. Schechtman. The Narrative Self. In *The Oxford Handbook of the Self*, ed. S. Gallagher, pp. 396-410. Oxford University Press, Oxford (2011).
- [27] D. Velleman. The Self as Narrator. In *Self to Self: Selected Essays*. Cambridge University Press, Cambridge (2006).
- [28] D. Dennett. The Self as a Center of Narrative Gravity. In *Self and Consciousness: Multiple Perspectives*, eds F. S. Kessel, P. M. Cole and D. L. Johnson. Lawrence Erlbaum Associates, Hillsdale, NJ (1992).
- [29] A. MacIntyre. *After Virtue: A Study in Moral Theory*. University of Notre Dame Press, Notre Dame, IN (1981).
- [30] G. Strawson. Against Narrativity. *Ratio*, 17:428-452 (2004).
- [31] R. Kane. *Ethics and the Quest for Wisdom*. Cambridge University Press, Cambridge (2010).
- [32] S. Kierkegaard. *Fear and Trembling and The Sickness Unto Death*. Trans. W. Lowrie. Anchor Books, New York (1954).
- [33] J. S. Mill. *On Liberty*.
- [34] J. M. Fischer, D. Pereboom and M. Vargas (2007). [Full citation details not specified in source document.]
- [35] C. E. Franklin (2013). [Full citation details not specified in source document.]

Fundamentalist Libertarianism and Emergentist Libertarianism

Daniele Conti

Abstract. Much of the credibility of libertarianism depends on whether it can explain, in a scientifically plausible way, how undetermined actions fit within a law-governed world. To tackle this task, libertarians have primarily pursued two distinct strategies: fundamentalist libertarianism and emergentist libertarianism. The first grounds agential indeterminism in the alleged indeterminacy of quantum mechanics, while the second holds that it arises from higher-level, emergent properties. It is often claimed that, of the two, the first is the only science-friendly option, whereas the second is dismissed as unscientific due to its incompatibility with the completeness of microphysical laws. In this paper, I contend that, on the contrary, the most scientifically plausible form of libertarianism is the emergentist variety.

After arguing that quantum-based libertarianism lacks empirical plausibility, the paper reassesses the standard evidential-paucity objection to emergentist libertarianism. I claim that this objection rests on a distorted interpretation of the available empirical evidence in physics – an interpretation often shared by emergentists as well – and an asymmetric allocation of the burden of proof. Using gravitation as a case study, I argue that current physics underdetermines fundamentalism and leaves open the possibility of genuinely emergent causal powers. The upshot is not that emergence is established, but that emergentist libertarianism is scientifically more plausible than typically assumed and cannot be dismissed on empirical grounds alone.

Chuck the cookbook: why instrumentalists should abandon formally-specified desert

Kristoffer Moody

Abstract. Are AI agents morally responsible? That is, are they justifiably subject to praise and blame for their actions? In this essay, I argue that if we understand moral responsibility as consisting in justified susceptibility to blame, then AI agents may be morally responsible.

It is often assumed that free will is valuable because moral responsibility is valuable, and is itself justificationally downstream of free will; if we lack free will, we lack moral responsibility—or so the thought goes.

This assumption is contested by semi-compatibilists, who claim that the standard challenges to free will (e.g., determinism) threaten free will, but not moral responsibility. This is because, as Strawson points out, our actual responsibility practices are not themselves predicated on our theoretical commitments: forming the theoretical belief that Jane lacks metaphysically-robust free-will is very unlikely to shield Jane from our indignation when she stomps on our feet.¹ Our practices are themselves relatively implastic; differing theoretical commitments do not have a strong effect on them. Still, our practices seem to stand in need of justification given how unpleasant it can be to be subject to them, e.g., when we are blamed.

An emerging strand of semi-compatibilist account identifies a desirable functional role for our reactive practices, and justifies those practices given that they serve that desirable functional role.² These *instrumentalist* accounts claim that blame functions to cultivate moral agency, and thus that blame is, in general, justified given that it serves that desirable end. These instrumentalist accounts typically assume that blame functions to cultivate the moral agency of those directly subject to that blame. They thus claim that justified blame requires that the target of that blame maintain certain capacities, e.g., control and poor quality of will. Absent these capacities, the thought is, blame would not serve to cultivate their moral agency; if Jane had been shoved, and this caused her to stomp on our feet, blaming her would not serve to cultivate her moral agency.

On the basis of these accounts, thus, for an agent to be a justified target of moralized blame requires that they maintain certain formally-specified capacities which would make them *deserving* of blame—i.e., they would maintain *desert*. Thus, these accounts claim that there is a formal *recipe* for justified blame which describes when agents maintain desert. Given these assumptions, it seems reasonable to presume that present AI agents do not maintain these capacities, and thus are not justifiably subject to blame.³

However, there is reason to be sceptical of the assumption that blame functions in this manner: empirical evidence from developmental psychology suggests that blame is, if anything, counter-productive for the children and adolescents subject to it.⁴ Further, blame seems to be a poor candidate explanation of the development of our moral agency even *were* it effective given the relative rarity with which we are blamed.⁵ However, we *witness* blame directed at *other agents* regularly, and we appear to rapidly internalize normative cues on the basis of such observation.⁶

Thus, I think it's reasonable to claim that blame's primary functional role is to *indirectly* cultivate moral agency in those who *witness* blame, rather than those directly subject to it.⁷ Given this assumption of blame's functional role, I claim that instrumentally-justified blame does not turn on agential features of the putatively blameworthy agent. Instead, I claim that blame is justified when an agent maintains *apparent controllability* over a bad effect. On this account, justified blame turns not on any capacities the putatively blameworthy agent may or may not possess, but instead on what they *appear* to possess. This is because, when it *appears* as though an agent had control over a given bad effect, blaming them serves to signal to observers that that behaviour is normatively unacceptable. Further, *not* blaming them would serve, in effect, to positively sanction that behaviour for observers. Thus, I will argue that instrumentalists should *chuck the cookbook* and abandon formally-specified desert.

This account of the conditions justifying blame is radically revisionary with respect to standard philosophical assumptions surrounding justified blame. However, I will argue that it is more conservative with respect to our actual practices than alternative instrumentalist accounts. This is because: (1) work on the moral psychology of blame appears to support a picture according to which blame is primarily predicated on apparent causal responsibility for bad effects.⁸ (2) Apparent controllability better explains the phenomenon of *resultant moral luck*, according to which identical agents performing identical actions may be differentially blamed depending upon the (un)lucky outcomes of their actions.⁹ (3) Assessments of control and quality of will are often *downstream* of blame in our actual practices; blame initiates a reactive exchange in which excuses may be offered. And (4) there is intriguing emerging experimental evidence that folk judgements of moral responsibility deny assumptions made by semi-compatibilist

⁴ Eisenberg et al. 2015, 123-124.

⁵ Shoemaker and Vargas, 2021, 590.

⁶ Nichols, 2021.

⁷ Moody and Nojournian, 2025.

⁸ Cushman, 2025.

⁹ Mazzocco et al., 2004.

¹ Strawson, 1963.

² Jefferson 2019; McGeer, 2019; Vargas, 2013.

³ Though see Vallor and Vierkant, 2024.

philosophers regarding the necessity of *control* for moral responsibility.¹⁰

If moral responsibility, understood as liability to praise and blame, is so understood as being in the eye of the beholder, then AI agents may possess all that is necessary for it: that they *appear* to have control over their actions.

REFERENCES

- Cushman, Fiery, 'Moral Judgment', in *The Handbook of Social Psychology*, ed. by D. T. Gilbert and others, 6th ed. (Situational Press, 2025).
- Eisenberg, Nancy, Natalie D. Eggum, and Tracy L. Spinrad, 'The Development of Prosocial Behavior', in *The Oxford Handbook of Prosocial Behavior*, ed. by David A. Schroeder and William G. Graziano (Oxford University Press, 2015), pp. 114–36, doi:[10.1093/oxfordhb/9780195399813.013.008](https://doi.org/10.1093/oxfordhb/9780195399813.013.008).
- Jefferson, Anneli, 'Instrumentalism about Moral Responsibility Revisited', *The Philosophical Quarterly*, 69.276 (2019), pp. 555–73, doi:[10.1093/pq/pqy062](https://doi.org/10.1093/pq/pqy062)
- Mazzocco, Philip J., Mark D. Alicke, and Teresa L. Davis, 'On the Robustness of Outcome Bias: No Constraint by Prior Culpability', *Basic & Applied Social Psychology*, 26.2/3 (2004), pp. 131–46, doi:[10.1207/s15324834basp2602&3_3](https://doi.org/10.1207/s15324834basp2602&3_3).
- McGeer, Victoria, 'Scaffolding Agency: A Proleptic Account of the Reactive Attitudes', *European Journal of Philosophy*, 27.2 (2019), pp. 301–23, doi:[10.1111/ejop.12408](https://doi.org/10.1111/ejop.12408).
- Moody, Kristoffer, and Makan Nojournian, 'Blame: What Is It Good For?', *Philosophical Explorations*, 28.1 (2025), pp. 32–50, doi:[10.1080/13869795.2024.2405523](https://doi.org/10.1080/13869795.2024.2405523).
- Murray, Samuel, and Thomas Nadelhoffer, 'Commonsense Morality and the Bearable Automaticity of Being', *Consciousness and Cognition*, 125.C (2024), p. 103748, doi:[10.1016/j.concog.2024.103748](https://doi.org/10.1016/j.concog.2024.103748).
- Nichols, Shaun, *Rational Rules: Towards a Theory of Moral Learning*, Oxford Scholarship Online, First edition. (University Press, 2021).
- Shoemaker, David, and Manuel Vargas, 'Moral Torch Fishing: A Signaling Theory of Blame', *Noûs* (HOBOKEN), 55.3 (2021), pp. 581–602, doi:[10.1111/nous.12316](https://doi.org/10.1111/nous.12316).
- Strawson, Peter, 'Freedom and Resentment', *Proceedings of the British Academy*, 48 (1963), pp. 187–211.
- Vallor, Shannon, and Tillmann Vierkant, 'Find the Gap: AI, Responsible Agency and Vulnerability', *Minds and Machines*, 34.3 (2024), doi:[10.1007/s11023-024-09674-0](https://doi.org/10.1007/s11023-024-09674-0)
- Vargas, Manuel, *Building Better Beings: A Theory of Moral Responsibility* (Oxford University Press, 2013).

¹⁰ Murray and Nadelhoffer, 2024.

The leeway–sourcehood distinction, the self, and AI

Tam Winter

Abstract. This paper offers a new interpretation of Kevin Timpe’s leeway–sourcehood distinction, arguing that it reflects a normative disagreement about which conception of the self is the rightful subject of freedom. On this reading, leeway-based conceptions of free will treat the empirical self—an agent’s uncriticised, everyday beliefs and desires—as the appropriate subject of freedom. By contrast, sourcehood conceptions ascribe value to the freedom of the authentic self, understood as a normatively privileged subset of an agent’s intentional mental states. To motivate these claims, I draw on the distinction between negative and positive freedom in political philosophy, arguing that it reflects a parallel normative disagreement over the rightful subject of freedom. By shedding light on the connection between conceptions of freedom and the self this suggests two promising directions for research into the conditions of free will for AI: whether AI is or could be capable of representing first-order beliefs and desires relative to counterfactual possibilities, and whether AI is capable of self-control, higher-order desires and volitions, and rationality.

1 INTRODUCTION

Kevin Timpe’s leeway–sourcehood distinction identified two different fundamental conceptions of free will in the contemporary literature: the leeway-based conception, and the sourcehood conception.¹ The leeway-based conception defines free will primarily in terms of the availability of alternative possibilities, typically understood as an agent’s freedom to do otherwise. By contrast, the sourcehood conception frames free will in terms of origination, requiring that the agent be the genuine source of their choices rather than factors external to them. While identifying a salient distinction in the literature, Timpe leaves the underlying source of this salience unexplained.

In this paper, I argue that the leeway–sourcehood distinction reflects differences of opinion about ‘the varieties of freedom worth wanting’ and that these are rooted in a deeper normative disagreement about which conception of the self is the rightful subject of freedom.² The leeway-based conception is best understood as taking the empirical self – the agent as we find them, with all their raw, uncriticised beliefs and desires – to be the rightful bearer of freedom. By contrast, the sourcehood conception ascribes value to the freedom of the authentic self, understood as a

normatively privileged subset of the agent’s beliefs and desires (e.g. those that are rational or reflectively endorsed).

To motivate these claims, I connect these different conceptions of the self to the negative–positive freedom distinction in political philosophy and argue that a similar normative disagreement underlies both the political freedom and free will debates. Identifying the common normative fault line running across political and metaphysical conceptions of freedom yields two main explanatory payoffs. First, it clarifies why the leeway–sourcehood distinction matters, helping to explain why some philosophers are drawn to one conception of free will rather than the other. Second, by shedding light on the connection between freedom and the self this analysis suggests two promising directions for research into the conditions of free will for AI: whether AI is or could be capable of representing first-order beliefs and desires relative to counterfactual possibilities, and whether AI is capable of self-control, higher-order desires and volitions, and rationality.

2 THE LEEWAY–SOURCEHOOD DISTINCTION

The leeway-based conception defines free will primarily in terms of the availability of genuinely open alternative possibilities, typically understood as an agent’s capacity to do otherwise. By contrast, the sourcehood conception frames free will in terms of ultimacy or origination, requiring that the agent have the capacity to be the origin of their choices rather than factors external to them. Some theorists emphasise leeway, others sourcehood, while many hybrid views combine elements of both. Timpe further distinguishes the leeway–sourcehood distinction from the question of whether free will is compatible with determinism by treating them as orthogonal, yielding four positions: leeway compatibilism, leeway incompatibilism, source compatibilism, and source incompatibilism.³ This framework provides a helpful map of the debate, but it raises further questions: why is this distinction so salient, and what is at stake between these competing conceptions? To address these questions, it is helpful to turn to another

¹ Timpe, 2005.

² Dennett, 1984.

³ Timpe, 2005, 12.

domain in which competing conceptions of freedom have been analysed: political philosophy.

3 THE NEGATIVE-POSITIVE FREEDOM DISTINCTION IN POLITICAL PHILOSOPHY

In political philosophy, a distinction is often drawn between negative and positive conceptions of freedom, as famously articulated by Isaiah Berlin in his 1958 lecture *Two Concepts of Liberty*.⁴ On the negative conception, freedom consists in a protected sphere within which a person may act as they wish, free from interference by others. In contrast, on the positive conception, freedom consists of a kind of self-mastery: the condition in which one's life is directed by one's true self rather than shaped by forces with which one does not identify. Freedom in these two senses can conceivably clash, for example if an agent's freedom from non-interference is curtailed by a smoking ban designed to promote freedom for their true self so that their actions are not determined by desires driven by addiction.

Building on Berlin's interpretation, George Crowder has argued that the negative-positive distinction principally reflects a normative disagreement about the subject to whom freedom ought to be ascribed.⁵ Negative conceptions treat the empirical self—the individual with all their existing beliefs and desires—as the rightful subject of freedom. Positive conceptions instead ascribe freedom to the authentic self, the part of an individual's personality associated with a normatively privileged subset of their beliefs and desires: for example, those that are rational or reflectively endorsed.

This analysis offers a plausible explanation of the connections between negative freedom and the absence of interference, and between positive freedom and the presence of self-mastery. When the subject of freedom is the empirical self, non-authentic desires are included as worthy of protection from constraint, rather than simply being seen as potential impediments to the authority of the authentic self. On the negative conception of freedom, what matters is whether the individual is prevented from doing what they want, even if their desires are irrational or inauthentic. Hence, their freedom is secured simply by the absence of obstacles such as interference by others. By contrast, when the subject of freedom is the authentic self, the absence of obstacles becomes less important than the presence of a particular kind of self-control. On the positive conception of freedom self-mastery is therefore

normatively central: the rational, authentic part of the individual must govern the whole if they are to be free.

4 REINTERPRETING THE LEEWAY-SOURCEHOOD DISTINCTION

This normative disagreement underlies the free will debate as well: the leeway-based conception is best understood as taking the empirical self to be the rightful bearer of freedom, while the sourcehood conception implicitly places weight on the freedom of the authentic self. Taking the leeway-based conception first, both compatibilist accounts such as AJ Ayer's and incompatibilist accounts such as Peter van Inwagen's frame free will in terms of the alternative possibilities required for the freedom to do otherwise.⁶ As Franklin has argued, free will understood in this way requires both that we have the ability and opportunity to do otherwise.⁷ We lack the opportunity to do otherwise only if there is some obstacle that prevents us from exercising our ability to do otherwise.⁸ Crucially, this conception of free will leaves aside the question of whether an agent's choices reflect their rational or authentic beliefs and desires or not. In this sense, freedom is conceived in a negative way: because no normatively privileged standpoint within the self is invoked, the central question is whether determinism is an obstacle that prevents the agent from doing otherwise than they actually do.

By contrast, the sourcehood conception defines free will in terms of ultimacy or origination. Advancing a hybrid but primarily source incompatibilist position, Robert Kane maintains that free will requires that the ultimate sources of our actions lie in us rather than in factors beyond our control, such as our genes or our upbringing.⁹ He frames free will not merely in terms of whether an agent could have done otherwise, but whether they could have done otherwise 'voluntarily (or willingly), intentionally, and rationally'.¹⁰ John Fischer defends a source compatibilist account that distinguishes between regulative control, which requires alternative possibilities, and guidance control, which does not.¹¹ For Fischer, freedom sufficient for moral responsibility consists in guidance control: acting through a reasons-responsive mechanism that the agent appropriately takes to be her own. Despite important differences, both accounts treat the capacity of an authentic, rational self to exercise self-mastery as normatively crucial to freedom, exhibiting a structurally positive framing. Free will consists not merely in the absence of obstacles, but in action and choice issuing from an appropriately governing standpoint within the self.

⁴ Berlin, 1998.

⁵ Crowder, 1988, 60

⁶ Ayer, 1954, 18. van Inwagen, 1983, 162.

⁷ Franklin, 2011, 692.

⁸ Franklin, 2011, 697.

⁹ Kane, 2007, 14

¹⁰ Kane, 2007, 20

¹¹ Fischer, 2005, 148.

5 THE EXPLANATORY PAYOFF

If this analysis is correct, there is a structural parallel between the negative-positive distinction and the leeway-sourcehood distinction: both are intimately connected to the empirical-authentic self distinction. Identifying this parallel has two main explanatory payoffs. First, it helps to explain what is at stake in the leeway-sourcehood distinction: a normative disagreement over which version of the self ought to be the subject of freedom. The leeway-based conception of free will attaches value to the freedom of the individual as we find them, with all their existing beliefs and desires, while the sourcehood conception places weight on the freedom of the authentic, rational or reflectively endorsed part of the individual's personality. These conceptions mark opposing poles of a spectrum, with most accounts combining elements of both to varying degrees. Even so, this normative disagreement remains a salient dimension of the free will debate and this analysis helps explain why some philosophers are drawn to one conception of free will rather than the other. Moreover, it suggests a deeper conceptual unity of freedom across metaphysics and political philosophy, indicating that debates about freedom and the self can be enriched by being studied across rather than solely within these domains. To illustrate this point, I will conclude by sharing some brief reflections about the implications of this analysis for research into the conditions of free will for AI.

6 IMPLICATIONS FOR AI RESEARCH

One of the central questions for the study of free will and AI is "what would it mean for AI to have free will in the sense required for moral responsibility?". If we hold the bar as high for non-human forms of intelligence as we do for humans, as I believe we should, then this question can be restated as: "what are the capacities required for the free will of intelligent agents?". The leeway-sourcehood distinction offers an approach to this question: requiring either that the agent, whether that is human or AI, have the capacity to do otherwise, or have the capacity for sourcehood.

But what is required for an agent to have the capacity to do otherwise or the capacity for sourcehood? For the former, the connection of leeway to the freedom of the empirical self suggests that the capacity to represent first-order beliefs and desires relative to counterfactual possibilities is central. For the latter, the connection of sourcehood to the freedom of the authentic self suggests that the capacity for self-control, higher-order desires and volitions, and rationality is of greater importance than counter-factual representation and action. To make progress on the

possibility of free will for AI, researchers could explore the extent to which AI systems have, or could have, such capacities. As many philosophers subscribe to a hybrid view of free will incorporating both leeway and sourcehood, a safe bet for future research into the free will of AI would be to pursue both lines of enquiry.

REFERENCES

- Ayer, A. J., *Philosophical Essays* (New York: St Martin's Press, 1954), pp. 3–20.
- Berlin, Isaiah, *The Proper Study of Mankind: An Anthology of Essays*, ed. by Henry Hardy and Roger Hausheer (London: Chatto & Windus, 1998).
- Crowder, George, 'Negative and Positive Liberty', *Political Science*, 40.2 (1988), pp. 57–73.
- Dennett, Daniel C., *Elbow Room: The Varieties of Free Will Worth Wanting* (Cambridge, MA: MIT Press, 1984).
- Franklin, Christopher, 'The Problem of Enhanced Control', *Australasian Journal of Philosophy*, 89.4 (2011), pp. 687–706.
- Fischer, John Martin, 'Reply: The Free Will Revolution', *Philosophical Explorations*, 8 (2005), pp. 145–156.
- Kane, Robert, 'Libertarianism', in *Four Views on Free Will*, ed. by Robert Kane, John Martin Fischer, Derk Pereboom, and Manuel Vargas (Malden, MA: Blackwell Publishing, 2007), pp. 5–43.
- Timpe, Kevin, 'Leeway vs. Sourcehood Conceptions of Free Will', *American Philosophical Quarterly*, 42.4 (2005), pp. 289–303.
- van Inwagen, Peter, *An Essay on Free Will* (Oxford: Clarendon Press, 1983).

What is a Prediction, Anyway? A Paradox of Everyday Agency and Scientific Prescience

Twan Kieboom¹

Abstract. This paper explores a paradox for hard-determinists with respect to the possibility of predicting the behaviour of an agent, human or otherwise, that is (1) informed of this prediction about itself, and (2) recalcitrant towards doing what it is predicted to do. This paradox has been the topic of philosophical discussion at least since MacKay (1960). We follow this discussion through Scriven (1965) to more recently Rummens and Cuypers (2010), incorporating Gijsbers' (2023) analysis in terms of Turing computability, and Rummens' (2024) subsequent reply. This paper argues that this literature has become burdened with unnecessary commitments: appeals to free choice or perfectly rational agents, external observers outside of the universe, and analogies between unpredictability and computational undecidability. This paper argues instead that there is a self-referential core to the paradox, which was already implicitly present in MacKay's original article, that allows for a straightforward solution on the basis of a conceptual analysis of the notion of prediction. Framed in terms of causal dependence and truthmaker theory, the crucial insight put forward here is that the paradox is resolved once we distinguish prediction from prescience. While not presenting an argument for libertarian free will, this analysis forces us to revisit the epistemology of the type of determinism that lies at the basis of incompatibilism.

1 INTRODUCTION

If strict determinism is true, then the initial state of the universe together with the laws of nature uniquely determines every subsequent state. This would entail that, in principle, everything is predictable: a sufficiently powerful intellect, given the initial conditions and the laws, could deduce any future event. This Laplacean picture has been challenged on various grounds: quantum indeterminacy, sensitive dependence on initial conditions, and computational complexity. But there is an even more fundamental challenge to this picture, which arises even if we were to grant, for the sake of argument, that we live in a universe that is perfectly deterministic and computable.

Consider the following scenario. A predictor announces that an agent will perform an action *A*. The agent, upon hearing this prediction, decides to do the opposite. Whatever the predictor predicts the agent will do, the agent will attempt to negate. This simple scenario, a predictor facing what Scriven (1965) called a 'contrapredictive agent', generates a genuine puzzle. Given that the agent is capable enough (or the predicted action sufficiently trivial), the predictor cannot succeed: if it predicts *A*, the agent does not-*A*; if it predicts not-*A*, the agent does *A*. It seems the agent can be unpredictable if it is the kind of agent that wants to be so. Yet, something will happen: either the agent will do *A*, or it will

not do *A*. As such, if the universe is deterministic by hypothesis and the agent is part of it, the agent's action should not be unpredictable.

The literature on this 'paradox of predictability' (Rummens and Cuypers 2010) has generated increasingly elaborate theoretical machinery, which this paper will revisit before putting forward its own resolution. The structure of this paper is as follows. In the next section we begin with how MacKay (1960) first captured the problem in terms of a 'logical indeterminacy'. In section three we then look at how Scriven (1965) offered a formal utilitarian explanation of contrapredictive motivation and a theorem of human unpredictability. In section four we discuss how Rummens and Cuypers (2010) distinguished 'embedded' from 'external' predictors and invoked an observer outside the causal order. In section five we see how Gijsbers (2023) connects the paradox to Turing's (1937) undecidability results and consider Rummens' (2024) reply, which questions the significance of these results for the paradox of predictability. Finally, in section six, the current paper's own thesis is put forward, which is that revealed predictions of contrapredictive agents necessarily fail due to a conflation of prediction and prescience.

2 MACKAY'S LOGICAL INDETERMINACY

MacKay (1960) argued against the familiar hard-determinist argument against free will: if our brains are physical systems governed by strict causal laws, then a sufficiently informed observer could predict our choices before we make them (MacKay 1960, p. 31). Despite his aim to carve out a compatibilism between freedom and determinism, MacKay's notion of free choice is ultimately incompatibilist: if our choices are absolutely knowable in advance, then they are not free. However, instead of arguing for a libertarian free will on the basis of quantum indeterminacy, he defends a 'logical indeterminacy' of the choice of an agent.

MacKay's response turns on a subtle point about the brain as "the organ of understanding" (MacKay 1960, p. 32). He argued that "where the brain-activity in formulating a statement is by itself sufficient to make it false and so out of date, an up-to-date formulation becomes impossible to the subject" (MacKay 1960, p. 37). He arrives at this conclusion by an argument based on the supposed identity of brain states and states of belief:

"Consider now the class of statements that describe the subject's own brain-processes or activities causally linked therewith. Among these, there will be a subclass which cannot be formulated or cannot be believed by the subject, without brain-activity that

¹ Dept. of Psychiatry, Amsterdam University Medical Research, The Netherlands. E-mail: t.kieboom@amsterdamumc.nl

interferes with the brain-process they describe.” (MacKay 1960, p. 37)

A true prediction about what an agent will choose, if communicated to the agent, must have already taken into account the effects of its being communicated to the agent’s brain. This rules out all those predictions carrying information that changes the neural activity of the agent in such a way that their subsequent action is changed as well. As I will explain below, according to MacKay, this opens up the possibility of a class of predictions which can be neither true nor false. This in turn gives rise to a paradox for the hard-determinist, who must assert that the truth of one or the other simply follows from cause and effect.

The key step in MacKay’s argument is the claim that for the agent, any proposition specifying their choice in advance is ‘logically indeterminate’ in this sense of being neither true nor false. Such predictions cannot reflect a truth about a state of affairs, because they cannot achieve universal validity. He reasons as follows: if a prediction P about an agent A would provide a truth, then both onlookers and A would always be right to believe P . But this is not the case for A , because the precondition of P ’s truth is that A cannot be influenced by P into acting such that P becomes false. From A ’s point of view, the action is still a matter of choice. This precondition is not fulfilled for the class of predictions about and communicated to the agent. The validity of such predictions, MacKay argues, is thus dependent on who believes them. While prediction P can in principle be true if only known to onlookers, it is fundamentally indeterminate for the predicted agent A , who would render it ‘out of date’ by getting to know it.

MacKay argues against the hard-determinist conception that free choice is an illusion stemming from a lack of predictive knowledge. If this were the case, the appearance of a free choice could be remedied by supplying an informative proposition that would relieve the agent of its ignorance about the truth value of prediction. However, if the agent were to entertain some proposition P describing what they will do, P would “lose its factual status” (MacKay 1960, p. 36). In effect, the brain activity involved in believing P would alter the very processes P describes. The onlookers therefore have no ‘predictive information’ they could give to the agent, no ‘prescience’ that may be shared: “(...) the place of the onlookers’ knowledge is already preoccupied for him [the agent] by the knowledge that the choice awaits his decision” (MacKay 1960, p. 37).

I believe that MacKay correctly identified the crucial structural feature of the paradox. If one accepts that all events are wholly determined by causal laws, then one must also accept that a prediction is itself an event that can causally influence what it predicts. If there is a causal pathway from prediction to predicted action, then the predicted system can in principle respond to the prediction in defiant ways—in which case the prediction *must* fail.

This insight is sound and important. However, MacKay might have overstated its implications. Consider for example his conclusion that “[f]or us as agents, any purported prediction of our normal choices as ‘certain’ is strictly *incredible*, and the key evidence for it *unformulable*” (MacKay 1960, p. 37, italics in original). Here he claims that communicated predictions are strictly indeterminate for predicted agents—that there is a fundamental perspectival relativity built into predictions about human actions that are communicated to the agent. But this overstates the case. Consider what Rummens and Cuypers (2010) call a “*fatalist mechanism* which always acts in accordance with a

revealed prediction” (Rummens and Cuypers 2010, p. 243). For example, if I am told I will choose A , and I am the sort of agent who accepts predictions about myself as authoritative, I will choose A . In that case revelation still occurs, but it does not defeat the prediction—it confirms it.

As such, MacKay (1960) might have been wrong that prediction of an agent’s actions is always indeterminate for that agent. Whether prediction succeeds or fails depends on how the agent responds to the prediction. Contrapredictivity, the supposed disposition to contradict revealed predictions, is essential to generating the paradox. Without further specifying the mechanism that could account for this disposition, MacKay’s (1960, p. 32) ‘relativity principle’ requires an unexplained capacity of ‘free choice’ on the side of the agent. The assumption of free choice directly contradicts the determinist assumption. Without the determinist assumption, there would be no paradox. What we need to address the paradox is thus (1) a plausible account of ‘contrapredictivity’ as well as (2) an account of why it is impossible for an otherwise perfectly prescient predictor to reveal predictions about a deterministic agent. We will first turn to contrapredictivity.

3 SCRIVEN’S UTILITARIAN ANALYSIS

Scriven (1965) puts the phenomenon of contrapredictivity on firmer footing. He starts from familiar cases of self-defeating predictions; opinion polls that, once published, change voting behaviour and thereby falsify themselves. Contrary to MacKay, such contrapredictive falsifications are not produced by the publication of predictions, but by “their comprehension and the assessment of their effect by listeners with a certain motivation and certain powers which brings about their defeat (perhaps unintentionally)” (Scriven 1965, p. 413, brackets in original).

What do these ‘certain motivations and certain powers’ consist of? Scriven defines ‘contrapredictivity’ quite precisely. For the sake of brevity I will not repeat his formalisms in full, but he held that an agent C is contrapredictive relative to a predictor P when the utility of falsifying P ’s predictions exceeds the calculated utility of not doing so (Scriven 1965, p. 414). He couched this notion in game-theoretic vocabulary about competitive situations where surprising one’s opponent carries value: “cards, business, love, and war” (p. 413).

In the same chapter, Scriven (1965, pp. 414–415) offered a formal theorem of unpredictability. Assume a rational predictor P inferring from information I_C the choice of individual C , where: (i) C is choosing rationally between alternatives and is physically capable of each; (ii) C is contrapredictive relative to P ; and (iii) C knows I_C is P ’s data and C has the computational capacity to derive whatever P would derive from it. The conclusion: I_C either leads to an incorrect prediction or none at all. C ’s choice cannot be rationally predicted by P from I_C .

The proof proceeds by contradiction. Suppose that P can successfully predict C . If I_C implies that C will choose action A , then C can derive this same implication (by condition iii). Being contrapredictive (condition ii), C will choose some different alternative, not- A . Hence the prediction is false, contradicting the assumption that P can successfully predict C . We should note here that, in this case, I_C apparently does not include C ’s contrapredictivity, and one could argue that as such P ’s information is simply incomplete, and we thus are merely dealing

with a case of inadequate data, which would seem to go against the very point of I_C .

Alternatively, if I_C includes as a fact that C is contrapredictive and also that C knows I_C is P 's data, then I_C cannot lead to any specific prediction in the first place, because P knows that C will predict P 's prediction and counteract it.

One should note here that at this point those defending determinism and those who believe in 'free' (contra)predictive agents may begin to accuse each other of begging the question. In the situation above both P and C are equally powerful predictors with complete information, but this just seems to point at an ill-defined notion of I_C . Because, as the determinist would argue, if there were real predictive knowledge to be had about C , this would just entail that C cannot act against predictions on the basis of this information (and the same would hold for P).

Indeed, if we assume I_C to be complete and true, instead of generating a paradox, conditions (ii) and (iii) of Scriven's theorem simply contradict each other. This is easily shown: if C is contrapredictive (ii), and capable of deriving whatever P derives on the basis of I_C (iii), then satisfying (ii) and (iii) would require P to derive the opposite of what it derives, because P 's complete and true I_C includes (ii) and (iii).

One result is that here we seem to be caught in a loop: either we assume some predictor P with access to some correct and complete I_C , in which case there cannot be a contrapredictive agent C , or the other way around. On its own, however, this does not yet establish the paradox. So far the possible existence of the (contra-)predictive agent is dependent on the non-existence of its counterpart. Of course, without having established that either may exist in the first place, these remain merely conflicting assumptions.

With respect to these predictors and their capacities, Rummens and Cuypers (2010) argue that Scriven "fails to make the crucial distinction between embedded and external predictability" (Rummens and Cuypers 2010, p. 240). They argue that the impossibility of P actually deducing the outcome does not imply that the available information is insufficient to logically infer that outcome. Instead, they argue that an observer not caught in the causal loop can still predict exactly what will happen. We turn to their argument now.

4 RUMMENS AND CUYPERS: EMBEDDED VERSUS EXTERNAL PREDICTABILITY

Half a century after MacKay (1960) and Scriven (1965), Rummens and Cuypers (2010) provide an analysis of the same problem that we have been discussing, which they coin the 'paradox of predictability'. They distinguish two notions of predictability. *External predictability* is defined as "the possibility of a (God-like) external observer, not part of the universe U , to make predictions of all the future events in U on the basis of its perfect knowledge of the initial conditions and the law-like function $f_U(U_0)$ " (Rummens and Cuypers 2010, p. 234). *Embedded predictability* holds that "if there exists a subsystem, S , embedded in U —for instance, a highly intelligent (human) being or a very

powerful computer—that is able to predict all the future events in U " (2010, p. 235).

Their key claim is that determinism implies external predictability but not embedded predictability. According to Rummens and Cuypers (2010, pp. 240-245) the paradox arises for embedded predictors because they face an overdetermined system of equations. I will repeat their argument here in shortened form. Let P be the predictor's prediction and A the agent's action. Both are physical events determined by the initial state of the universe U_0 , according to law-like functions g and h :

$$\begin{aligned} P &= g(U_0) \\ A &= h(U_0) \end{aligned}$$

The predictive task requires the solution $P = A$. But when revelation occurs and the agent is contrapredictive, we also have $A = \text{not-}P$. Substituting:

$$P = \text{not-}P$$

This is unsolvable. The embedded predictor cannot fulfil the predictive task. So far, this is very similar to the result of our discussion of Scriven (1965) and MacKay (1960).

But Rummens and Cuypers' argument continues: they introduce, in addition to the predictors discussed in the literature so far—all of which were embedded predictors—the notion of the external predictor. An external predictor can make a prediction P^* that is not itself a physical event in the universe. P^* is not subject to the contrapredictive condition (it is not the case that $A = \text{not-}P^*$). As such, the external predictor can correctly predict that the embedded predictor will output some value (say, $P = 0$) and that the agent will therefore do the opposite ($A = 1$). The external predictor's task, $P^* = A$, is perfectly solvable: $P^* = 1$.

The analysis by Rummens and Cuypers is clarifying in several respects. They identify the source of the paradox in the self-referential structure created when predictions causally influence what they intend to predict.² With the notion of an external predictor who is not caught in the causal loop, they make plausible the idea that determinism does not need to imply a logical indeterminacy. It only implies an unsolvable equation for an embedded predictor. As such, they specify MacKay's (1960) idea that a sufficiently 'secretive' predictor can still perfectly predict the outcome.

Contrary to MacKay (1960), Rummens and Cuypers (2010) believe that "(embedded) unpredictability is neither a sufficient nor a necessary condition for human freedom" (p. 246). They do however introduce a notion they call "take-it-or-leave-it-control" (p. 245), which grants informed agents the power to go against predictions about themselves. Of course, for hard-determinists such "take-it-or-leave-it control" would still beg the question much like Scriven's (1965) "contrapredictivity" that we discussed in the previous section. The determinist's point is that *if* it can be determined that A will happen, *then* there shouldn't be such a thing as take-it-or-leave-it control with respect to A (see also Gijsbers, p. 582). But Rummens and Cuypers (2010) explicitly intend to account for this. As we saw above, they state that A , including its perceived take-it-or-leave-it control, is fully determined by this deterministic function $h(U_0)$. Their argument is just that this

² Ismael (2019) is very much in alignment with Rummens and Cuypers (2010) on this point as well. She writes: "From a logical perspective the problem arises because even though it is missing no information, a predictor can't generate a correct prediction because the nomological

determinants of the predicted event includes the prediction itself. That means that in order to make a correct prediction, the predictor has to predict what it will predict. As soon as it predicts what it will predict, it has to update its prediction on that information" (p. 490).

function could only be known by a predictor external to the universe.

Yet, we may ask ourselves what we are to make of this appeal to such an ‘external’ observer? While remaining “neutral as to the intrinsic nature of this external standpoint” (Rummens and Cuypers 2010, p. 234, footnote 2), they describe the external observer as “God-like” and “non-physical” (p. 234), making predictions that are “not part of the law-like causal chain of events in the deterministic universe” (p. 233). Indeed, if defined as such, the external predictor can still satisfy hard-determinism for actions happening within the universe. But what does it mean to conclude this from the standpoint of embedded agents like ourselves? Is it supposed to be able to make determinism true in principle, but just not for us? If anything, this obscures the actual paradoxical situation that we find ourselves in as embedded agents: that apparently we have good reason to believe that we can resist certain communicated predictions of our everyday actions, if we so wish it.

It would seem that Rummens and Cuypers’ (2010) argument, while saving determinism in the abstract, would thus still invalidate the hard-determinist position for any concrete situation. Even situations in which a communicated (i.e., embedded) prediction could rely on fatalistic mechanisms to come true (coercion, agreement, direct brain intervention, etc.), no hard-determinist could ever succeed in showing that some action A is *strictly* determined by some function $h(U_0)$, without severely tampering with our normal everyday experience. After all: $h(U_0)$ would not just predict that you won’t jump off a bridge (even if you are told you would not), it would also predict things like that you will not clap your hands while on that bridge (and this you might very well do, even if just for the sake of argument). It would seem, to me at least, that this still amounts to MacKay’s original conclusion that predictive knowledge about the agent’s actions remains ‘logically indeterminate’ whenever it is accessible by the agent. We will return to this in the final section.

For now, leaving aside for a moment the metaphysical status of the external predictor and the implication for the debate between compatibilists and incompatibilists, may we at least conclude that Rummens and Cuypers’ (2010) distinction between embedded and external predictors has solved the paradox? Let me explain why I think this might not be so. Consider that, next to being outside of the causal chain, the only key characteristic of the external predictor P^* is that it is in possession of the law-like function that *actually* determines A ’s response to P ’s prediction. Let’s call this function $g^*(U_0)$, such that $P^* = g^*(U_0)$. Given contrapredictivity, we know that this function must be different from whatever function P is calculating, because, as Rummens and Cuypers (2010) have shown, P ’s function cannot produce the correct result.

This presents a paradox of its own. Because whatever operations the function $g^*(U_0)$ specifies, the underlying assumption by Rummens and Cuypers remains that this is a

function that is in principle computable. At the very least this would imply that it is a well-formed mathematical expression that could in principle be written down, and therefore be physically represented in our universe. In principle, then, we could reduce any external predictor to an embedded predictor. This clearly reinstates the paradox: it is not so much that an embedded predictor cannot correctly predict the action of a contrapredictive agent, but rather that the *correct* mathematical function describing the evolution of the universe cannot—and for reasons unknown—be revealed, not unlike a grandfather that cannot be killed by its time-travelling grandson.

Similarly, Gijsbers (2023) responds to Rummens and Cuypers (2010) that “[i]f the external predictor EX goes through some process of reasoning, then it is in principle possible to have an embedded predictor EM that goes through exactly the same process” (p. 585). In his reply to Gijsbers, Rummens (2024) seems to grant the point that the embedded predictor would essentially be able to run any algorithm that the external predictor might run, but dismisses the relevance of this result: “Crucially, however, we have never assumed in our example that the action A is not decidable (‘predictable’) in this logical sense for EM” (Rummens 2024, p. 2099).

However, this reply does not seem to consider an important implication of setting up the paradox in terms of computable functions. That is to say: if the external predictor is computing the *correct* function, then this function is supposed to be *causally precluded* from being expressed in a universe where a contrapredictive agent A would have access to it. Without such causal preclusion, any accessible function used to predict A could simply not be the *correct* function. But such a causal preclusion argument does not chime well with the idea that the evolution of the universe is governed by some law-like function that takes in initial conditions and evolves these according to physical laws. There is no physical law that precludes physical laws from being expressed mathematically. If we are to allow for such a causal preclusion here, we might as well allow for the causal preclusion of contrapredictivity in the first place, in which case there would be no paradox to begin with. The force of the paradox consists precisely in the intuition that the fatalistic mechanisms that would be required for such causal preclusions are simply not present in our everyday lives.

One might object that it would be impossible to find out what function P^* is computing, given that we do not have any causal connection to it.³ This would work if the external predictor truly was not constrained in any way whatsoever (see also Gijsbers, p. 589). Rummens and Cuypers’ assumption is, however, that the external ‘Laplacian’ predictor is still constrained with respect to doing something akin to producing well-formed mathematical expressions that capture the deterministic evolution of the universe. For my argument to work the only thing that is required

³ Consider for example Ismael (2019), who presents her own solution to the paradox of prediction on the basis of the light cone structure of Minkowski space-time. I take her argument to entail that a function like $h(U_0)$ could never be possessed by an embedded predictor, or ‘natural oracle’ as she calls it, because special relativity fundamentally does not allow for the simultaneous representation in the here-and-now of all events in space-time that are to determine an event in the future. That is to say: from within the system there is no way to acquire the information required to specify U_0 . This solution defines determinism in a way that no longer

accommodates a Laplacian predictor, however does allow that “(...) the future of the universe as a whole is completely determined by the state along any complete spacelike hypersurface” (Ismael, 2019, p. 489). This argument thus seems to target not so much the fundamental possibility of construing some function $h(U_0)$, were we to guess the correct initial conditions, but only the fundamental possibility of verifying such a guess. But the paradox obtains just as well for unverifiable guesses of the correct deterministic equation.

is the sheer theoretical possibility of a lucky conjecture of such a well-formed expression.⁴

5 GIJSBERS AND COMPUTATIONAL UNDECIDABILITY

Gijsbers (2023) offers a diagnosis of the paradox of predictability that is somewhat different from that of the authors previously discussed. He argues that the paradox of predictability is structurally identical to Turing's proof that the halting problem is undecidable.

Let us first quickly recap the non-formal version of Turing's (1937) famous proof. We start by assuming a decision procedure H that correctly determines, for any program and input, whether that program will halt after a finite number of steps, or loops forever. Next, we construct an algorithm G that takes as its input a program and queries H 's prediction about whether the program halts when given itself as input. But then G does the opposite: if H says the program will halt, G loops forever; if H says the program will loop, G halts. Finally, we apply G to itself, such that if H predicts for input G that program G will halt, G loops, and vice versa. As such, whatever $H(G, G)$ predicts, it is wrong. As a consequence H cannot exist, and thus the halting problem is undecidable.

Gijsbers draws two conclusions from this analogy. First, he states that "any feature of the real world that does not have an analogon in Turing's proof must be inessential to the derivation of unpredictability" (2023, p. 590). According to him, the paradox is fundamentally about undecidability, not about being embedded in the physical universe. Second, he argues this shows that Rummens and Cuyper's (2010) distinction between embedded and external predictors, as they formulated it, cannot be correct. An external predictor using algorithm H faces the same problem as an embedded predictor if H is embodied in the universe.

In his reply to Gijsbers (2023), Rummens (2024) basically distinguishes these two claims as two versions of Gijsbers' thesis. The *strong undecidability thesis* claims that all examples of the paradox are rooted in undecidability. The *weak undecidability thesis* claims only that some examples are. Rummens accepts the weak thesis by admitting the "artificial" (Rummens 2024, p. 2098) possibility of an embedded predictor simulating an external predictor, but rejects the strong thesis.

Against the strong thesis, Rummens argues that Gijsbers misunderstands their original example. According to him, the situation, which we described in section four, is not at all undecidable. In fact, anyone could understand what will happen. That is to say: the embedded predictor knows perfectly well that if it outputs $P = 0$, the action will be $A = 1$. Note that P here does not need to be a prediction about anything special; it can be about something as mundane as predicting that A will not raise a cup,

and A raising the cup just to make the point that they *can* raise a cup as a response to being told they will not do it (or vice versa). The paradox is that the function which the embedded predictor uses to determine the evolution of the universe, should also determine whether or not A raises their cup. Again, the problem is that the predictor "cannot reveal this correct prediction without thereby invalidating it" (Rummens 2024, p. 2100).

Indeed, from our perspective of knowing the exact computational rule, or 'causal law', which governs the contrapredictive agent, the situation is *fully decidable*. What is impossible for the predictor is therefore not the successful simulation of the contrapredictive agent's *computation* of an output from an input, but rather the *formulation of its true input*. When faced with the contrapredictive agent, the predictor cannot successfully formulate a prediction that constitutes a statement of knowledge about the future (a statement of 'prescience').

The difference between embedded and external predictability is not computational but *causal*. Both embedded and external predictors can *compute* the correct answer when given the correct input; only the external predictor can 'output' it as a description of a future state-of-affairs without that statement triggering the contrapredictive response. For the embedded predictor, knowing that any output will change what *should* have been its input, knowledge of the workings of the contrapredictive agent is simply useless.

A similar diagnosis applies to the weak thesis, which Rummens accepts. Gijsbers' Turing-machine construction does generate a genuine instance of undecidability of prediction: the universal predictor cannot exist. But the only thing this undecidability shows amounts to a point I have already made in section four without an appeal to Turing's undecidability proof: if a computable function describing the universe's evolution is in principle physically representable, then the embedded/external distinction cannot insulate the external predictor from the paradox.

Rummens (2024) concludes his reply by stating that "[i]f we want to see the paradox of predictability in action" we should "not be side-tracked by formal results about undecidability" (p. 2103). At the same time, however, as we concluded above, the external observer presents its own side-track as well.

6 WHAT IS A PREDICTION, ANYWAY?

What does the paradox look like without reference to such an external predictor? We may note that Rummens (2024) does at some point summarise his position about what lies at the root of the paradox without a reference to the external predictor:

"As soon as people (or machines)—secretly or openly—try to use their knowledge of the future to

⁴ While it is, of course, not a serious possibility to do something akin to guessing the exact initial conditions of the universe, this is not a theoretical problem. After all, the assumption behind the paradox of predictability is already that of a Laplacean predictor who has perfect knowledge of initial conditions. However, what *is* a theoretical problem is the possibility that the information that needs to be computed to make actual predictions with the external predictor's function cannot be sufficiently compressed to be expressed in the embedded agent's universe. In other words: if the universe is *computationally irreducible* (see for example Wolfram 2002, pp. 737-750), then it would not be possible to predict its evolution without wholly simulating it. This is the type of argument Rummens (2024) is

looking for when he states that the relevant question is whether an "external algorithm exists that successfully predicts specific or—possibly all—events in that universe" (p. 2102). But in my mind, allowing the external predictor to pick an algorithm that is not physically instantiable for the embedded agent just amounts to constructing a fatalistic mechanism to make a prediction come true. We have already granted that relying on fatalistic mechanisms remains possible for embedded predictors too, so I don't see how this would reinstate the relevance of the embedded/external distinction.

have a causal effect on the world, their supposed knowledge becomes vulnerable to counterpredictive invalidation by an appropriate reasons-responsive mechanism.” (Rummens 2024, p. 2103)

The interesting thing that happens here is the switch from ‘knowledge’ to ‘supposed knowledge’. Indeed, this I believe, is the root of the paradox. Either we have ‘knowledge of the future’, in which case our predictions should be impervious, or—by contraposition—our predictions are fallible, in which case we do not have such knowledge.

MacKay (1960), Rummens and Cuypers (2010), and Rummens (2024) believe that we do have the right kind of ‘counterpredictive control’ (Rummens 2024) to falsify predictions about ourselves. However, in the previous sections we saw that the notion of the contrapredictive agent begs the question for the hard-determinist. That being said, I believe Jenann Ismael (2019) shows very clearly that the capacities of a contrapredictive agent do not have to amount to anything controversial. She conceives of a ‘counterpredictive device’ which just takes as input a prediction about its own output, and then outputs the opposite (e.g., changing a 1 into a 0). Even if we do not fully understand the mechanism by which we ourselves (or our brains) are capable of performing such an operation, we do understand how we could program this contrapredictive mechanism into a very simple computing device. Note that such a device does not assume the abilities of a (perfectly) rational predictor, as Scriven (1965) had to, even though it would allow us to exercise counterpredictive control in our everyday lives. As such, I believe the paradox still stands. The philosophically important conclusion we should draw from this, I believe, is one that is not so much about free will or determinism, as it is about *prediction* and *prescience*.

So, what is a prediction, anyway? I believe this becomes clearer once we look at predictions through the lens of truthmaker theory (see Armstrong 2004). The basic idea here is that propositions are true or false in virtue of things in the world called truthmakers. Now, a prediction P is a proposition that purports to have future state-of-affairs A as its truthmaker. P claims to represent A . But as far as P is concerned, A does not exist yet, and thus cannot provide the basis for acquiring knowledge of P ’s truth.⁵ Now, because the future follows from the present (and does so independently of there being causal laws that determine *how* the future follows upon the present), A can in principle be caused by P , but certainly not the other way around. The root of the paradox is thus that predictions fundamentally get the direction of dependence wrong. The fact that this only shows in those particular (and peculiar) situations where we do not find the world and each other so constrained as to conform to our representations of their respective futures, does not warrant the move from prediction to prescience.

Let me contrast this to the conclusion that Gijsbers (2023) arrives at:

“[W]e must reject the idea that determinism implies predictability, and reject this on a combination of formal and material grounds: Turing’s proof on the formal side, the possibility of embedding prediction processes into our universe on the material side. The

laws of nature may determine that B will follow from A, without knowledge of those laws giving us an effective algorithm for computing or deducing B from A.” (Gijsbers 2023, p. 591)

The crucial difference here between my account and that of Gijsbers (2023) is that I believe determinism *does* imply predictability: a representation of the future derived from the *present* state-of-affairs on the basis of a conjectured “iron chain of cause and effect” (MacKay 1960, p. 33). The determinism that is supposed to threaten everyday agency just *is* the idea that there is something like Rummens and Cuypers’ (2010) function $A = h(U_0)$, which would determine everyone’s fate. What determinism does not imply, however, is *prescience*: a representation of the future derived from the *future* state-of-affairs. If the universe allows for a function $h(U_0)$, then the paradox entails that we cannot have a solution to this function in our possession.

Indeed, like Turing’s undecidability proof, this argument proceeds on logical grounds only. It does not explain where or how this hypothetical h breaks down, it just shows it cannot exist *and* be revealed to a contrapredictive agent. Deterministic predictions that have to take into account their own output can simply turn out to be self-defeating. The difference between the undecidability of the halting problem and that of the paradox of prediction is that, whereas Turing shows that his h cannot exist, the paradox of prediction only shows that a contrapredictive agent cannot have access to h . This becomes apparent in the special case when our future is presented as a given, whereas we can only act as if it is a choice on the basis of the presentation.

7 CONCLUSION

The paradox of predictability, properly understood, is not a puzzle about free will or computational limits. It is a puzzle about *what prediction is*. A prediction purports to be a representation of a future state-of-affairs. For the prediction to succeed, the future state must match the current representation. MacKay (1960), Scriven (1965), Rummens and Cuypers (2010), Gijsbers (2023), and Rummens (2024), all point to this being impossible for a special kind of prediction that allows for counterpredictive control. This is because in those cases the representation itself has a causal effect on the future state by which the prediction becomes self-defeating.

Indeed, this was MacKay’s insight when he wrote that “where the brain-activity in formulating a statement is by itself sufficient to make it false and so out of date, an up-to-date formulation becomes impossible” (1960, p. 37). What he described as the ‘logical indeterminacy of a free choice’ is better understood as a conflation of prediction and prescience. Such a prediction about a contrapredictive agent does not fail because the future is necessarily indeterminate due to some form of spontaneous agent causation; it fails because the prediction cannot occupy the epistemic position it claims to occupy—that of a *fact* about what *will* happen, instead of a *link* in the *causal chain*. Such a prediction is itself among the causes of the very future state-of-affairs which it would require to ground its own truth.

⁵ Even if we were to take an ‘eternalist’ position in truthmaker theory, i.e., the idea that true claims about the future are ‘eternally’ true or false in virtue of the actual future state that will obtain (or has always obtained on a block universe conception), this does not matter for our argument. To be

able to claim a prescience about the future, we would still be required to either have some reliable access to this future state, or be able to guess it correctly. The paradox of prediction, however, shows precisely that there cannot be such access, nor such a guess.

Of course, it is possible to constrain the contrapredictively inclined agents so as to be unable to negate the prediction. If we define, as Scriven (1965) did, the contrapredictive agent as having the required capacities to resist any prediction in principle, this ultimately begs the question whether an agent so defined actually exists (see also Gijsbers 2023, p. 582). But few would reject the idea that in very mundane situations people can negate low-stakes predictions about themselves.

What matters to the distinction between prediction and prescience in general is whether the prediction enters the causal chain leading to the predicted event. If you can keep it from doing so then prediction does amount to prescience. In such cases, however, the predictor is more like an experimenter or manipulator, isolating its subject or victim from information relevant to the latter's decision making. This does not so much show that the subject has no agency as it does that in such situations it is the predictor who is in control. At best this implies a shift in agency rather than show that it does not exist.

It is illustrative here to consider Daniel Dennett's (2012) scenario in which a neurosurgeon—today we could perhaps imagine an AI—would claim to control your will:

“[I]f our neurosurgeon tells you she controls your will, you can easily refute her claim decisively—if you are not overwhelmed by the authority of her white lab coat and fancy apparatus—and, of course, if she did prove right, you'd better hope some friends of yours can remove you from her clutches. But until that imaginable, logically possible but extraordinarily unlikely state of neurotechnology is achieved, you can be quite sure that your free will is not an illusion at all” (p. 28).

Dennett considers here how scientific or indeed 'deterministic' knowledge can be used to take on the role, not of a *predictor*, but of a *manipulator* with the capacities to make its predictions come true. This is exactly the real threat we face as agents. Our agency is not in danger of being predicted, but in danger of being circumvented. And this is what we should resist.

We may conclude that the implications of the paradox of predictability are more epistemological than metaphysical. We are tempted to think of knowledge of the future as analogous to knowledge of the past or present: there is a fact of the matter, and we either represent it correctly or not. But the paradox shows that this analogy breaks down for a specific class of cases. Whenever our representation of the future causally contributes to bringing about the actual future, we cannot coherently claim this representation equates to *knowledge* of what will happen. What will happen may still depend on our *act* of making the predictive claim.

This is therefore not a refutation of determinism and predictability per se, but it does qualify the notion's usefulness for our everyday agentive behaviour. What is refuted is the idea that predictors—whether they be human or artificial intelligences—could acquire perfect prescience, without first convincing everyone of fatalism. All predictors relevant to our everyday lives are embedded predictors. Such predictors might describe their predictions as if we were detached from them, but we must know that we are not.

REFERENCES

- Armstrong, D. M. (2004). *Truth and Truthmakers*. Cambridge: Cambridge University Press.
- Dennett, D. C. (2012). *Erasmus: Sometimes a Spin Doctor is Right*. Amsterdam: Praemium Erasmianum Foundation. Available at: <https://dl.tufts.edu/pdfviewer/s4655t074/xw42nm33z>
- Gijsbers, V. (2023). The paradox of predictability. *Erkenntnis*, 88, 579–596.
- Ismael, J. (2019). Determinism, Counterpredictive Devices, and the Impossibility of Laplacean Intelligences. *The Monist*, 102(4), 478–498.
- MacKay, D. M. (1960). On the logical indeterminacy of a free choice. *Mind*, 69, 31–40.
- Rummens, S. (2024). The roots of the paradox of predictability: A reply to Gijsbers. *Erkenntnis*, 89, 2097–2104.
- Rummens, S., & Cuypers, S. E. (2010). Determinism and the paradox of predictability. *Erkenntnis*, 72, 233–249.
- Scriven, M. (1965). An essential unpredictability in human behaviour. In B. B. Wolman & E. Nagel (Eds.), *Scientific Psychology: Principles and Approaches* (pp. 411–425). Basic Books.
- Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1), 230–265.
- Wolfram, S. (2002). *A New Kind of Science*. Champaign, IL: Wolfram Media.

On moral responsibility, free will, and intentional agency in AI systems

Paulius Rimkevičius

Abstract. Can artificial intelligence (AI) systems have free will? In a recent paper, which draws on earlier work by Daniel Dennett (2003), Christian List (2025) has argued that they can. In that paper, he defines free will in terms of three conditions. First, the entity should be capable of intentional agency. Second, alternative possibilities should be open to that entity. Third, the entity should have causal control over its actions. He then argues that there is no reason why AI systems could not, in principle, have these capacities, and sketches some reasons for thinking that current AI systems do indeed have at least some of them. When discussing intentional agency, he focuses on evidence that suggests that AI systems can have the intentional states of belief and desire. He only indirectly addresses the question whether they can have intentions, or how to distinguish intentional and unintentional actions in the case of AI systems. However, as David Chalmers (2025) has recently noted, whether AI systems can have intentions, or can act intentionally or unintentionally, is a question with its own complications, distinct from those related to belief and desire.

In this paper, I therefore consider some ways to further develop List's proposal by specifying what it would take for an AI system to have an intention or to act intentionally. With List's definition of free will in mind, I will review recent attempts to formally define intentions and intentional actions, specifically for AI systems (Ashton 2023; Córdoba et al. 2023; Judson et al. 2024a, 2024b; Ward et al. 2024; Williams 2025). I will argue that these definitions (especially the one by Ward et al. 2024), are largely in line with List's proposal. However, they are also more minimal than those that are dominant in philosophy and neuroscience of intentions and intentional actions (see Yaffe 2022; Parés-Pujolràs & Haggard 2022). For example, the formal definitions leave no role for consciousness, while the dominant definition in neuroscience presupposes consciousness (Parés-Pujolràs & Haggard 2022). Therefore, both List's proposal and these formal definitions are open to the charge that they change the topic. Importantly, these authors are explicitly committed to capturing our ordinary concepts. So, if the formal definitions capture something else than the ordinary concept, it would count as a drawback, at least according to their own criteria.

However, most claims in the philosophical and neuroscientific literature about the ordinary concepts of intention and intentional action have been speculative. So, secondly, I will turn to review recent empirical work on these ordinary concepts, in order to evaluate these claims (Kozuch & Nichols 2011; McGuire 2016; Deutschländer et al. 2017; Nolte et al. 2020; Vierkant et al. 2020; Vekony et al. 2020; Sytsma et al. 2022). I will argue that empirical research on the ordinary concepts of intention and intentional action suggests that it is likewise more minimal than the above-mentioned dominant definitions in philosophy and neuroscience.

If so, then the charge against List's proposal, and the attempts at formalisation, has less strength than one might have thought, at least from the perspective of List himself and those who take ordinary concepts to be relevant. Additionally, the fact that our ordinary concepts of intention and intentional action are quite minimal can help to explain why, as recent empirical work suggests, ordinary people are inclined to attribute free will and moral responsibility to AI systems (Voiklis et al. 2016; Stuart & Kneer 2021; Astobiza 2024; Kneer & Markus 2024), without also suggesting that they overattribute a rich psychological concept (Halina & Shevlin 2019).

In sum, according to List, AI systems can have free will, because they can have intentional agency, alternative possibilities, and causal control. While List does not go into much detail on how we should understand intentional agency here, his proposal is in line with recent formal definitions of intentions and intentional actions for AI systems. However, like those definitions, List's proposal is open to the charge that it changes the topic to something more minimal, compared to the dominant definitions in philosophy and neuroscience, and our ordinary concepts. A closer look at the empirical evidence suggests, however, that our ordinary concepts are similarly quite minimal. This also helps to explain the evidence that suggests that ordinary people would indeed attribute free will and moral responsibility to AI systems. I will finish the paper by briefly considering the implications of this conclusion for other influential views of free will and moral responsibility, beyond those of Dennett and List.

REFERENCES

- Ashton, H. 2023. Definitions of intent suitable for algorithms. *Artificial Intelligence and Law*, 31: 515–546. <https://doi.org/10.1007/s10506-022-09322-x>
- Astobiza, A. M. 2024. Do people believe that machines have minds and free will? Empirical evidence on mind perception and autonomy in machines. *AI and Ethics*, 4: 1175–1183. <https://doi.org/10.1007/s43681-023-00317-1>
- Dennett, D. C. 2003. *Freedom Evolves*. London: Penguin.
- Deutschländer, R., Pauen, M., Haynes, J.-D. 2017. Probing folk-psychology: do Libet-style experiments reflect folk intuitions about free action? *Consciousness and Cognition*, 48: 232–245. <https://doi.org/10.1016/j.concog.2016.11.004>
- Chalmers, D. J. 2025. Propositional interpretability in artificial intelligence. <https://arxiv.org/abs/2501.15740>

- Córdoba, F. C., Judson, S., Antonopoulos, T., Bjørner, K., Shoemaker, N., Shapiro, S. J., Piskac, R., Könighofer, B. 2023. Analyzing intentional behavior in autonomous agents under uncertainty. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 42: 372–381. <https://doi.org/10.24963/ijcai.2023/42>
- Halina, M., & Shevlin, H. (2019). Apply rich psychological terms in AI with care. *Nature Machine Intelligence*, 1: 165–167. <https://doi.org/10.17863/CAM.37897>
- Judson, S., Elacqua, M., Córdoba, F.C., Antonopoulos, T., Könighofer, B., Shapiro, S. J., Piskac, R. 2024a. “Put the car on the stand”: SMT-based oracles for investigating decisions. In *ACM Symposium on Computer Science and Law (CSLAW '24)*, 73–85. <https://doi.org/10.1145/3614407.3643699>
- Judson, S., Elacqua, M., Cano, F., Antonopoulos, T., Könighofer, B., Shapiro, S. J., Piskac, R. 2024b. soid: a tool for legal accountability for automated decision making. *Proceedings of Computer Aided Verification: 36th International Conference*, 233–246. https://doi.org/10.1007/978-3-031-65630-9_12
- Kneer, M., Christen M. 2024. Responsibility gaps and retributive dispositions: evidence from the US, Japan and Germany. *Science and Engineering Ethics*, 30: 51. <https://doi.org/10.1007/s11948-024-00509-w>
- Kozuch, B., Nichols, S. 2011. Awareness of unawareness: folk psychology and introspective transparency. *Journal of Consciousness Studies*, 18(11–12): 11–12. <https://www.ingentaconnect.com/contentone/imp/jcs/2011/00000018/f0020011/art00006;jsessionid=57va2ufuc4wqu.x-ic-live-01#trendmd-suggestions>
- List, C. 2025. Can AI systems have free will? *Synthese*, 206: 115. <https://doi.org/10.1007/s11229-025-05209-x>
- McGuire, J. 2016. Can one decide to do something without forming an intention to do it? *Analysis*, 76(3): 269–278. <https://doi.org/10.1093/analys/anw036>
- Nolte, A., Buckwalter, W., Rose, D., Turri, J. 2020. Deciding without intending. *Journal of Cognition*, 3(1): 1–14. <https://doi.org/10.5334/joc.101>
- Parés-Pujolràs, E., Haggard, P. 2022. What are intentional actions? In Maoz, U., Sinnott-Armstrong, W. (eds.). *Free Will: Philosophers and Neuroscientists in Conversation*. Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780197572153.003.0021>
- Stuart, M. T., & Kneer, M. (2021). Guilty artificial minds: folk attributions of *mens rea* and culpability to artificially intelligent agents. *Proceedings of the ACM on Human-Computer Interaction*, 5 (CSCW2): 1–27. <https://doi.org/10.1145/34795>
- Sytsma, J., Bishop, R., Schwenkler, J. 2022. Has the side-effect effect been cancelled? (No, not yet.). *Synthese*, 200(5): 1–15. <https://doi.org/10.1007/s11229-022-03864-y>
- Vekony, R., Mele, A., Rose, D. 2020. Intentional action without knowledge. *Synthese*, 199: 1231–43. <https://doi.org/10.1007/s11229-020-02783-0>
- Vierkant, T., Deutschländer, R., Sinnott-Armstrong, W., Haynes J.-D. 2019. Responsibility without freedom? Folk judgements about deliberate actions. *Frontiers in Psychology*, 10: 1133. <https://doi.org/10.3389/fpsyg.2019.01133>
- Voiklis, J., Kim, B., Cusimano, C., Malle, B. F. 2016. Moral judgments of human vs. robot agents. *25th IEEE international symposium on robot and human interactive communication (RO-MAN)*: 775–780. <https://doi.org/10.1109/ROMAN.2016.77452>
- Ward, F. R., MacDermott, M., Belardinelli, F., Toni, F., Everitt, T. 2024. The reasons agents act: intention and instrumental goals. *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024)*: 1901–1909. <https://dl.acm.org/doi/10.5555/3635637.3663053>
- Williams, I. 2025. Intention-like representations in language models? <https://philarchive.org/rec/WILIRI-4>
- Yaffe, G. 2022. What is an intention? In Maoz, U., Sinnott-Armstrong, W. (eds.). *Free Will: Philosophers and Neuroscientists in Conversation*. Oxford: OUP. <https://doi.org/10.1093/oso/9780197572153.003.0001>

Graded Minds, Graded Agency

Mehmet Mirioğlu

Abstract. Can machines have free will? This paper advances a conditional thesis: a philosopher who consistently endorses panpsychism cannot exclude artificial intelligence from the domain of free will. The argument proceeds in three stages. First, a dual-aspect neutral monist framework is developed, grounding the claim that every physical process has a mental counterpart and that mental richness scales with physical complexity. A transitivity argument shows that the absence of sharp thresholds in evolutionary and developmental history makes any sudden emergence of experience metaphysically implausible. Second, the components of free will are analysed within a compatibilist framework. Drawing on Fischer and Ravizza's account of reasons-responsiveness, the paper introduces two original moves: an Object Argument that reduces the ownership condition to the question of objecthood, and a recharacterisation of reasons-responsiveness in terms of its physical correlates, namely integration capacity and output repertoire flexibility. Both components are shown to be graded rather than threshold properties. Third, this framework is applied to artificial intelligence. A mechanical analogue of the biological spectrum is constructed and AI systems are evaluated against the established criteria. The paper concludes that attributing at least minimal free will to AI is not optional for the consistent panpsychist; withholding it would be an inconsistency the framework cannot sustain.

1. INTRODUCTION

Since Turing's 1950 paper, the question "can a machine think?" has been one of the most intensely debated questions in philosophy. Attempted to be answered either through a behavioral criterion or a biological substance condition, this question has been reignited with the advent of contemporary artificial intelligence systems. Yet, there remains another question, cast in the shadow of these debates and almost never asked: are machines free? Just as much as whether a system thinks, whether its actions genuinely belong to it, whether it can act in a reason-responsive manner, and consequently, whether it possesses free will, deserves to be questioned. This article addresses precisely this question. The central claim I will defend in this article is as follows: anyone who consistently adopts panpsychism cannot avoid attributing a minimal level of free will to artificial intelligence. This is a conditional thesis and depends on the acceptance of panpsychism. The goal is not to convince the reader of panpsychism, but rather to demonstrate to someone who leans toward this position that artificial intelligence cannot be excluded from the debate on free will. In doing so, it will first be shown why panpsychism ought to be taken seriously, and the framework of dual-aspect

monism will be established as the most consistent form of this view. Within this framework, it will be argued that every physical interaction has a mental dimension, and that this dimension enriches alongside physical complexity. Subsequently, the components of free will will be examined, demonstrating that being an agent is related to being an object, whereas reason-responsiveness is related to a physical data-processing capacity. After establishing that all these components possess a gradient structure, the final section will discuss the position of artificial intelligence within this spectrum and substantiate the main claim.

2. THE PANPSYCHISM CONDITION

This article conditions the claim of attributing a minimal level of free will to artificial intelligence upon the thesis of panpsychism. This is a conditional thesis: rather than focusing on what an approach to consciousness that rejects panpsychism would posit, we are concerned with why a consistent panpsychist would have to be more open-minded regarding artificial intelligence. If consciousness is part of the fundamental reality of the universe, one should also be more willing to attribute free will to artificial intelligence. At first glance, this approach may seem counterintuitive. Taking a position that is already highly controversial in the philosophy of mind as a foundational condition will admittedly appeal to a niche audience. Consequently, the condition I propose might resemble a counterfactual along the lines of "if elephants were pink, humans could fly." However, despite our general intuition dictating otherwise, I argue that this condition rests upon an ever-expanding and strengthening literature. Although it is not the dominant view at present, I believe it will be taken more seriously over time. The purpose of this article is not to convince the reader of panpsychism, but rather to initiate a debate among philosophers who already lean toward this position and to make a modest contribution to the growing literature in this field. Nevertheless, I am aware that the condition involves a "big if." Therefore, instead of attempting to prove panpsychism, I will try to convey at least the intuition and motivation that "this approach might not be so absurd after all."

Today, the view that an experiential dimension exists in every physical element of the universe is not widespread. However, this has not always been the case throughout history. Numerous philosophers whose reasoning skills

we can firmly trust have looked favorably upon this view. Thales attributed the magnet's attraction of iron to its possession of a soul¹. Heraclitus acknowledged the existence of the Logos, the living, rational principle that permeates the cosmos and governs all things². The Stoics conceived of the cosmos through a universal breath as a single, massive, rational, and living organism³. Bruno argued that every particle in the universe has an inner life⁴, while Spinoza maintained that the universe possesses a mental dimension just as it does a physical one⁵. Leibniz constructed the universe as a structure composed of atom-like, perceptual centers of energy, asserting that even a stone is an aggregate of sleeping monads⁶. Schopenhauer viewed everything from gravity to human desire as a manifestation of the cosmic will⁷, and Russell developed neutral monism⁸. James, Whitehead, Bacon, and Cavendish also offered interpretations of the mind that are compatible with the fundamental reality of the universe⁹. This idea, which may seem marginal today, was never deemed "absurdity" in the history of philosophy; on the contrary, it has been a remarkably common intellectual intuition. Moreover, this prevalence manifests itself not only in academic philosophy but also in well-established theological traditions. The hierarchy of being in traditional theology, extending from inanimate objects to plants, humans, and angels, does not signify a transition from the absence of consciousness to its presence, but rather a gradual intensification of divine consciousness and perception. In Biblical texts, the mountains shouting for joy¹⁰, all creation groaning as in the pains of childbirth¹¹, and the declaration that if humans fall silent, the very stones will cry out the truth¹², imply that nature is not a blind and unfeeling mechanism. The Qur'an explicitly ascribes a hidden awareness to every element of the universe by declaring that the seven heavens and the earth glorify Allah in their own tongues¹³, and that the heavens and the earth responded to the divine address by saying, "We come willingly."¹⁴ The intuition that consciousness is a constitutive element of the universe lies at the common denominator of both philosophical thought and deeply rooted theological texts. This, of course, does not constitute a proof. However, it indicates that the conditional statement of this article ought to be approached with greater permissiveness.

The second part of the motivation toward panpsychism stems from the impasse in which rival positions find themselves. Dualist and physicalist approaches either face severe difficulties in explaining the nature of consciousness or grapple with irresolvable questions regarding the nature of the interaction between the physical and mental domains. For hundreds of years, dualism, and for nearly a century, physicalism, have continued to reckon with the fundamental problems of the philosophy of mind. While one side takes sophisticated steps to explain how mutually irreducible properties interact, the other side fails to answer the question of how qualitatively distinct dimensions can be reduced to one another. In this predicament, there is no harm in approaching dual-aspect monism and panpsychism more warmly once again. Moreover, panpsychism can be positioned not as a rival to these approaches, but as a complementary framework. Strawson positions consciousness as the intrinsic and hidden nature of physical matter, demonstrating that a realistic physicalism necessarily entails panpsychism¹⁵. Goff fills the ontological space left vacant by contemporary physics with mathematical formulas through panpsychism, defending the proposition that "the inside of the physical is consciousness."¹⁶ Chalmers, on the other hand, proposes a solution to dualism's interaction problem by spreading consciousness everywhere in the universe like an irreducible law of nature¹⁷. These three examples demonstrate that panpsychism can both fill the qualitative gap physicalism falls into while explaining matter, and alleviate dualism's causal problem. In this sense, panpsychism functions as a unifying framework.

2.1. Dual-Aspect Neutral Monism

I will ground dual-aspect neutral monism, which claims that the universe has both a physical and a mental aspect and that these are aspects of a substance that is neither purely physical nor purely mental, through a sort of transitivity argument. To understand this argument, it is necessary to make a distinction between primitive properties and emergent properties. Emergent properties manifest themselves through the combination of the universe's primitive properties in specific ways. While no single H₂O molecule is liquid on its own, the collective interaction of these molecules can bring about the property of liquidity. Properties that do not exist in the parts can manifest in the whole. However, it should be noted that both liquidity and the interactions between H₂O molecules are physical; physical interactions have thus brought about new physical properties. Primitive properties, on the other hand, are properties that are intrinsic to the structure of the universe, do not derive from other things, lie at the foundation of emergent properties, and possess a boolean structure. For example, physicality is a primitive property of this universe. The mass of a cell, the coordination capacity of an organism,

¹ Aristoteles, *De Anima* 405a19-21 in Skrbina 2005, p. 25

² Diogenes Laertios IX: 5-12 in Skrbina 2005, p. 31

³ Sambursky 1959, p. 36 in Skrbina 2005, p. 65.

⁴ Hoeffding 1908, pp. 124-126 in Skrbina 2005, p. 87

⁵ Spinoza 1677/1994.

⁶ Leibniz 1714/1991.

⁷ Schopenhauer 1818/1958.

⁸ Russell 1921.

⁹ Other important thinkers who could be included in this tradition include Henri Bergson, Charles Sanders Peirce, Charles Hartshorne, Pierre Teilhard de Chardin, Gustav Fechner, and Rudolf Hermann Lotze. For a broader overview, see Skrbina 2005.

¹⁰ Psalm 98:8.

¹¹ Romans 8:22.

¹² Luke 19:40.

¹³ Al-Isra 44.

¹⁴ Fussulat 11.

¹⁵ Strawson 2006.

¹⁶ Goff 2017.

¹⁷ Chalmers 2013.

or the connection density of a brain involves accumulation and graduality. But "being physical," unlike these properties, either exists or does not. No entity transitions suddenly from being non-physical to physical; such a metaphysical leap appears absurd. Physicality is not something gained when complexity exceeds a certain threshold, but a fundamental property that everything that exists has carried from the very beginning. The claim put forward by this article is that structurally the same expectation holds true for subjective experience. Cognitive skills such as language, abstract thought, and theory of mind develop and are acquired gradually. But the "capacity to experience things" that lies at the foundation of these emergent properties (namely, the existence of qualia) either exists or does not, just like physicality. The number of things an entity can experience may increase, its experiential repertoire may expand, and it may possess more complex experiential networks. Yet, just like physicality, experience either exists or does not. This binary structure of qualia is the basis for the personal and evolutionary continuity arguments that will be used to demonstrate that consciousness is the fundamental reality of the universe.

First, the transitivity within our personal history would require the primitive forms of consciousness to be traced back to a single cell. To understand this, let us imagine an album consisting of photographs taken every single day. This album can be seen as a biographical document spanning from a person's birth to their adulthood. Each photograph is almost identical to the previous one: the same body, the same nervous system, the same behavioral pattern; it carries only very minor, cumulative differences. Is there any photograph in this album, then, that provides a good reason for us to think that while experiencing nothing the day before, this entity began to experience things the very next day? While the entire physical system, all behavioral outputs, and the entire neurological organization remain almost identical to the previous day, how could a mysterious effect have occurred, causing this entity to now possess a subjective experience? It is impossible to answer this question without appealing to a Cartesian substance dualism. Moreover, accepting that such a leap exists also lays the groundwork for a radical skepticism regarding the minds of others. If the ability to experience emerges through such sharp transitions, it becomes impossible for us to know whether any other entity has crossed that threshold. The difference between the previous day and the next day is microscopic every single time. Nevertheless, if an entity that experienced nothing the day before while displaying identical behaviors suddenly possesses experience the next day with almost nothing having changed, there remains no reason to think that we are not currently among such philosophical zombies. Furthermore, this regression can be sustained for every single day. Along the line extending from adult to child, and from child to infant, cognitive skills gradually regress. The infant lacks a theory of mind, has not acquired language, and cannot think abstractly. But these are properties of cognitive

complexity; they are not properties of the existence of qualia. We have no justification for denying that an infant feels pain or experiences satiety or discomfort. On the contrary, neonatal research reveals that infants' pain responses exhibit neurological patterns comparable to those of adults¹⁸. Well, if the infant feels something, why should the moment of birth be the boundary for the existence of this feeling? How could the severing of the umbilical cord or the departure from the amniotic fluid initiate subjective experience? There is no reason to think that these physical conditions have a deterministic effect on qualia. Consequently, the same argument will continue in the womb: if the fetus a second before birth has qualia, it also has them a second before that. And a second before that. There is no point at which to halt this regression. Embryo, blastocyst, zygote... At each step, the difference between adjacent moments is minimal, and at no step is there a mechanism that justifies the transition from zero experience to experience. Therefore, proto-qualia, no matter how minimal and different from ours they may be, must be traceable all the way back to the zygote. The zygote is a single cell. This means that the property of being able to experience must also exist at lower levels than complex living beings. Since the question of how this property of the zygote emerges from the movement of atoms cannot remain unanswered, this property must be traceable all the way down to the fundamental structure of the universe.

When we extend the photograph album analogy to the evolutionary axis, the same conclusion will naturally manifest itself. Let us imagine a massive album extending from you to your father, your grandfather, and across the entire lineage of ancestors, gradually reaching hairy, small-brained primitive primates, fish, and single-celled organisms. Again, each photograph is remarkably similar to the one before and the one after it. Cognitive capacities gradually diminish in this album: first language, then abstract thought and imagining the future, then tool-making, and finally social learning. Yet, at each step, one can ask whether this entity experiences things or not. Just as in personal history, it is accepted in evolutionary history that, except for extreme cases, there are no sudden leaps. If the individual in one photograph possesses the property of being able to experience, the ancestor from the previous generation, who is almost identical to them, must also possess this property. Consequently, even though cognitive properties increase and decrease cumulatively, the tracing of qualia must extend all the way back to single-celled organisms. Considering that single-celled organisms are also the products of millions of years of chemical evolution, the same transitivity can be traced all the way down to the fundamental structure of the universe.

The first critique that comes to mind is that this argument might be a classic form of the sorites paradox. When you remove grains of sand one by one from a heap of sand,

¹⁸ See Goksan et al. 2015.

you cannot find a single point where removing one grain causes it to cease being a heap. Nevertheless, certain collections of sand grains do not constitute a heap. One might think that the argument relies on this faulty reasoning. However, the distinction made at the beginning of the section between primitive and emergent properties is necessary precisely to prevent this error. "Being a heap of sand" is an emergent property, just like having language skills, thinking abstractly, or solving complex problems. Emergent properties can be formed through a transitive process via the accumulation of primitive properties. If I had based my reasoning on the human capacity to think and concluded that the zygote could also think, the sorites critique would be accurate. But the argument is built upon the premise that primitive properties cannot manifest themselves through a metaphysical leap. That is, it operates like an argument stating that because a heap of sand is physical, a single grain of sand must also be physical. The physicality of a grain of sand does not fully replicate the physicality of a heap of sand, yet the grain of sand is still physical. The experience of a zygote does not fully replicate the experience of a human, yet the zygote still possesses experience. The accumulation of sand grains can bring about the heap of sand and its associated emergent physical properties in a gradient manner; the transitivity argument does not hold for these emergent properties. The same applies to the situation where entities possessing experience manifest complex cognitive properties. The transitivity argument cannot be applied to these cognitive properties. The property of physicality is binary, while its richness is continuous. The property of experientiality is also binary, while its richness is continuous. The fact that consciousness is gradient and rests upon a spectrum does not undermine its underlying boolean nature.

The most consistent explanation of this transitive process is found within the framework of dual-aspect monism, which argues that the universe has both a physical and a mental aspect. First, it must be conceded that mental and physical properties appear distinct at first glance. None of the properties such as what it is like to experience something, namely phenomenality; the fact that mental states are always about something, namely intentionality; the fact that what pain feels like can only be known from a first-person perspective, namely third-person inaccessibility; and the mind's capacity to generate rules of logic, morality, and rationality and aim to comply with them, namely normativity, appear to be properties of physical objects. On the other hand, the mind and body do not seem independent of each other either: it can be observed that the desire to drink water activates muscles, brain damage alters character, and morphine alleviates pain. This situation gives rise to the classic mind-body problem: how can these two domains coexist when they are, on the one hand, mutually irreducible, and yet, on the other hand, seem to interact? Dual-aspect monism proposes the following to resolve this tension: there is no real causal relationship between the physical and the

mental; they are two different and simultaneous aspects of the same reality. Although it is thought that the desire to drink water moves the muscles, what actually occurs can be described as follows: an increase in blood sodium levels activates the hypothalamus, signals reach the prefrontal cortex, the premotor cortex, basal ganglia, and primary motor cortex are stimulated sequentially, acetylcholine is released, and muscle fibers contract. While all of this is happening, the feeling of thirst, the desire to drink water, selective attention, the intention to move, and kinesthetic awareness are experienced simultaneously. The exact same process can be described entirely in a physical language or entirely in a mental language; both narratives express the same event from two different perspectives. This relationship can be likened to two sides of a coin: when you flip a coin, both the heads side and the tails side move simultaneously. Not because one causes the other, but because they are two aspects of the same reality. For this reason, while the complete explainability of physical processes within themselves is called the principle of physical closure, I propose to term the capacity of mental processes to form a complete and consistent causal chain within themselves as the principle of mental closure. The same reality can be expressed through two separate causal processes that are closed within themselves. This reality itself is neither physical nor mental: the physical and the mental are appearances of an underlying neutral reality from two perspectives; as envisioned by Russell's tradition of neutral monism, both are two distinct accounts of something more fundamental. In this sense, dual-aspect monism is neither materialism, idealism, nor substance dualism: there is a single reality, but this reality can be explained from neither a purely physical nor a purely mental perspective. And if this framework is correct, the conclusion reached by the transitivity argument developed above becomes highly intelligible. As physical complexity increases, mental richness will also increase, because the physical and the mental are two inseparable sides of the same reality; one cannot speak of the total absence of a mental dimension in any physical interaction. In the rest of the paper, we will proceed from this metaphysical framework to analyze the components required for free will.

3. THE COMPONENTS OF FREE WILL

If we are speaking of the free will of artificial intelligence, it is important to clarify what we mean by this. The foundation of this article will be based on the compatibilist approach to free will and will proceed from its definition regarding reason-responsiveness. But first, let us clarify our intuition.

The most common intuition regarding free will is that a person can be held responsible only when they could have acted otherwise. For instance, we think that a bank teller who hands over money at gunpoint does not act freely, because they have no real alternative option. Conversely, we think that someone who makes a voluntary donation

acts freely, because they could have chosen not to give. Similarly, thinking that someone who raises their hand by reflex is not free, while someone who raises their hand despite being able to act otherwise is free, proceeds from a similar foundation. The philosophical equivalent of this intuition is called the principle of alternate possibilities¹⁹: an agent possesses free will only when they could have acted otherwise. This principle has played a decisive role in both libertarian and certain compatibilist accounts of free will. According to the libertarian view, genuine free will requires a capacity that falls outside of causal determinism, where the agent can genuinely choose between alternate paths. In this view, free will and causal determinism are incompatible; if the universe possesses a deterministic structure, no one has the possibility to do otherwise, and therefore no one is genuinely free.

However, Frankfurt counterexamples can be used to demonstrate the deficiency of this principle and, naturally, of libertarian free will. Indeed, according to these counterexamples, it is possible to deem a person as possessing free will even if no alternate option exists, provided that the appropriate conditions are met. For instance²⁰, Jones decides to kill Smith entirely of his own free will and without any external pressure, and executes this action. However, there is a circumstance of which Jones is unaware: a device implanted in Jones's brain will activate the moment Jones inclines to abandon his decision, forcing him through a neural intervention to execute this action. Because Jones decides to kill Smith on his own, the device never activates. Events unfold stemming entirely from Jones's own mechanism. Yet, it is also true that if Jones had attempted to act otherwise, the device would have prevented it. That is to say, Jones was never genuinely in a position to do otherwise. According to the principle of alternate possibilities, Jones should not be held responsible. However, our intuition says exactly the opposite. It seems correct to hold Jones responsible for this action, because the action emerged from his own independent mechanism, no intervention took place, and the device never activated. The presence of the device belonged merely to a hypothetical scenario and had no effect on the actual events. This case demonstrates that moral responsibility and free will stem not from being able to do otherwise, but from how the action comes about. The principle of alternate possibilities is not a necessary condition for free will. The genuine foundation of free will should be sought not in the agent's access to alternate scenarios, but in the structure of the mechanism from which the action emerges.

Well, if alternate possibilities are not required, what is required for free will? The most consistent answer to this question is provided by criteria based on reason-responsiveness. According to this view developed by Fischer and Ravizza, free will depends on two conditions. First, the action must emerge from a reason-responsive

mechanism²¹. A reason-responsive mechanism is a mechanism that possesses the capacity to react differently in the face of a sufficient reason. An agent acting under hypnosis or through brain manipulation is not reason-responsive; because no matter how powerful a reason is presented, that mechanism will continue to operate in the exact same manner. Conversely, an agent exercising normal practical reasoning is reason-responsive; when confronted with a sufficiently powerful reason, they will act differently. Second, that mechanism must be the agent's own mechanism²². This is the ownership condition. Even if a device implanted externally into someone's brain is reason-responsive, that mechanism does not belong to the agent and is therefore not sufficient for free will. Together, both conditions demonstrate that the agent possesses guidance control over their action

This approach makes free will possible even in a deterministic world. Causal determinism does not contradict whether the agent's mechanism is reason-responsive or not. A system can be both causally determined and reason-responsive. Because reason-responsiveness asserts that if a sufficient reason had been presented, the agent acting through the same type of mechanism could have reacted differently. That is, even if one is bound to act in that manner under those specific conditions, the ability to act differently under different conditions will render the individual free. Consequently, the view in question is compatibilist: free will can be seen as compatible with causal determinism.

In this article, we will interpret these two conditions within the framework of dual-aspect monism. I had claimed in dual-aspect monism that every physical process has a mental counterpart, and vice versa. In the remainder of the article, I will attempt to analyze the physical aspect of these two components necessary for free will. First, I will re-evaluate the condition based on being an agent, ownership, and possessing a mind, and subsequently, reason-responsiveness within this framework. Afterwards, these components will lead us to clearer criteria of examination regarding whether AI has free will or not.

3.1. The Object Argument

If we wish to attribute free will to something, that thing must first be an agent. If a reason-responsive mechanism lacks a mental and personal dimension, it would not be meaningful to say that it possesses free will. The fundamental characteristic of being an agent can be evaluated as possessing a perspective. The "ownership" criterion, which Fischer and Ravizza count as a criterion for reason-responsiveness, is also related to this. For determining whether a mechanism belongs to that person first requires determining who that person is. If it is known to whom the perspective belongs, it can then be

¹⁹ Frankfurt 1969, p. 829.

²⁰ For the original example, see Frankfurt 1969, s. 831-833.

²¹ Fischer and Ravizza 1998, pp. 38-48.

²² Fischer and Ravizza 1998, pp. 207-210.

determined whether the operating mechanism also belongs to that person. Of course, since proto-perspectives come together to form more compact perspectives within the panpsychist framework, the question "what happens such that small perspectives come together and form a distinct agent?" will inevitably arise. This is a combination problem that panpsychism must overcome, and it is necessary to reach a conviction regarding this to detail the debate on free will. Let us recall that within the framework of dual-aspect monism, every mental event has a physical aspect, and every physical event has a mental aspect. Since the mental aspect of the universe is inaccessible to third parties, the method for answering mental questions will consist of finding and examining the physical counterpart of this mental aspect. The physical counterpart of this question is the special composition question in metaphysics: What happens such that objects come together to form compact objects? Because the physical counterpart of being an agent is being an object. Whenever small objects come together to form a complex object, proto-perspectives come together at that exact moment to form a new perspective. Therefore, what it means to be an object must first be investigated. In short, understanding what kind of entities are agents depends on understanding what kind of entities are objects. This question is the subject of this section.

Let us begin by addressing the two extreme answers among the most fundamental responses given to this question. The first is atomism: it defends that the only real objects in the universe are fundamental particles, and that nothing else constitutes an object in a genuine sense. According to this view, there is actually no such thing as a table, there are only atoms arranged tablewise; there is no such thing as a galaxy, there are only clusters of atoms. Calling a table or a galaxy an object is merely a matter of practical convenience. The second extreme answer defends the exact opposite: there is only one genuine object in the universe, and that is the universe itself. Everything else consists merely of names we assign to parts of the universe for practical reasons. Although these two views proceed in opposite directions, they arrive at a common conclusion: both reject the reality of what we ordinarily call objects. However, it is possible to reject both extreme answers with a single observation. I am neither a fundamental particle nor the entirety of the universe. Nevertheless, I cannot doubt that I am an object. Even to question whether I am an object requires me to exist as an object, to be an integrated system conducting that inquiry; a thought that rejects my objecthood implicitly assumes from the outset the existence of the object carrying that thought. Therefore, there is at least one entity in the universe that is not an atom yet is an object: myself. And there is at least one entity in the universe that is not the entirety of the universe yet is an object: again, myself. This simultaneously refutes both atomism and the view that the universe is the sole object. There are real objects in the universe, neither at the level

of the smallest particle nor at the level of the whole, but in between them, at the medium scale.

The first thought that might come to mind here is that the criteria for being an object are met only by biological entities, and therefore, only biological organisms fall into the category of real objects²³. I find this objection justified to some extent; there may be unique qualities that distinguish biology from other physical interactions. However, being biological is itself a transitive property without a sharp boundary. When chemical evolution is considered, biology does not emerge all of a sudden; it develops gradually. As a matter of fact, there are structures whose biological status is still debated today. Viruses, viroids, and prions possess some characteristics of life while lacking others; they are considered neither fully animate nor fully inanimate. Furthermore, there is no harm in claiming that biological structures themselves are formed by the gradual accumulation of non-biological chemical processes. That is to say, even if there was a first cell, its origins in evolutionary history must have been non-biological chemical processes. Naturally, the transitivity process on which biological evolution relies, discussed in the previous section, can also be adapted from chemical processes to biological processes. The history of the earth owes its biological past to chemical processes that do not yet deserve to be called biological. If we grant objecthood only to biological entities, we would have to invent a bizarre category like "neither an object nor a non-object" for these borderline structures. Moreover, biology is ultimately a product of physical interactions. If physical interactions can come together to form a biological object, what makes biology privileged in this regard? There seems to be no obstacle for a mechanical system to be an object as long as it meets the same criteria.

The conclusion we have reached so far is that the criteria for being an object are not unique to me or to biological entities. However, since biology is our only concrete example, we must arrive at these criteria by proceeding from biology and then generalize them to other entities. The criteria emerging from the biological example are that my parts exist in a tight unity with one another, interact causally in a systematic manner, and maintain their internal order within certain boundaries. In this regard, integration, systematic causal interaction, and self-regulation stand out as the fundamental criteria that make something an object. My purpose here is not to provide a flawless list of the necessary and sufficient conditions of objecthood; my central claim is that any structure that satisfies these criteria is an object and therefore possesses a perspective. At the very least, if you are persuaded by dual-aspect monism. An important feature of these criteria is that they are not binary, but involve degrees. The more integrated and self-regulating a system is, the stronger of an object it will be. While a bacterium satisfies these criteria at a minimal level, the

²³ van Inwagen 1990, p. 82

colonies of *Volvox*, *Pandorina*, and *Eudorina*, which represent the transition from single-celled to multicellular organisms, demonstrate how gradual this process is. If we consider a bacterium an object, we must also consider a cell in my body an object, because the conditions that make the bacterium an object hold true for my cell as well. Yet, those cells come together to form a larger object; this means that an object can contain other objects within itself, and each new integrated structure forms a new object, and consequently, a new perspective and a new agent. (This can be grounded by the fact that single-celled organisms separated from the colony can continue to live.) That is to say, the sub-object does not disappear; a new structure is superposed upon it, which is entirely consistent with the idea of gradient agency. The dual-aspect framework states that every object is accompanied by a mental dimension. Whatever turns a totality of physical interactions into a single object, the accompanying mental dimension constitutes the perspective of that object. Therefore, a bacterium or a computer, to the extent that it is an object, possesses a perspective. By virtue of the principle of equal treatment, artificial intelligence systems must also be objects to the extent that they satisfy the required criteria of integration and systematic causal interaction.

We will discuss the position of mechanical systems within this spectrum in the next section; however, before examining whether artificial intelligence satisfies these criteria, we must analyze the other component of free will.

3.2. Reason-Responsiveness

Let us recall the Frankfurt counterexamples once more. What is the difference between Jones who commits murder because of the chip and Jones who commits murder by his own will? The willful Jones would not have killed Smith if he had been provided with a different reason, for instance, if he had been told that Smith was innocent. However, the chipped Jones would have killed Smith no matter what reason was provided. The genuine difference here appears to be not the existence of alternate possibilities, but the reason-responsiveness of the mechanism that drives Jones to action. For even if a person is bound to act in that specific manner under those precise conditions and reasons, that is, even if determinism operates, there is no obstacle to calling them free-willed as long as they can act differently under different reasons. Proceeding from this intuition, Fischer and Ravizza's proposal is that the decisive criterion for free will is the reason-responsiveness of the mechanism possessed by an agent. According to them, reason-responsiveness has two components: receptivity²⁴ and reactivity²⁵. Reason receptivity is the system's capacity to recognize reasons, grasp how they relate to one another, and exhibit an understandable pattern across a wide spectrum, including moral reasons. Reason reactivity, on the other hand, is the capacity to actually act differently in

the face of a sufficient reason in at least one possible scenario, consistently with the laws of the actual world. Neither of these two components needs to be satisfied in an extremely robust manner; for even humans who are ordinarily deemed responsible occasionally ignore sufficient reasons. Yet, a very weak responsiveness is not sufficient either; reacting to a single reason by mere coincidence cannot be the bearer of moral responsibility. Therefore, a moderate level is sought: a regular and understandable receptivity accompanied by at least a weak level of reactivity. The exact criteria of this moderate reason-responsiveness and its differences from the strong and weak versions require a much more detailed examination; since these details exceed the main scope of this article, I will confine myself here to presenting only the fundamental framework.

Although receptivity and reactivity are good starting points, these concepts appear highly laden with mental terms. However, as stated previously, since mental concepts are not suitable for evaluation from a third-person perspective, it is necessary to find their physical counterparts. First of all, a reason is a sort of data-processing method, and its physical counterpart can be said to be related to input-output systems through a functionalist interpretation. Input-output systems are also gradient by nature. For instance, a stone, roughly speaking, can only process data regarding gravity and specific electromagnetic inputs. A magnet processes electromagnetic data across a wider spectrum in addition to gravity. A bacterium, beyond these, possesses the capacity to process chemical inputs. In the process extending from worms to humans, the data-processing process develops slowly and gradually. Yet, the concept of a reason should not be considered as consisting merely of an input-output system. For example, receptivity involves the integration, comparison, and weighting of signals coming from different sources beyond mere input transmission. Integration, which is one of the criteria for being an object, manifests itself here as well. The richer the integration a nervous system exhibits, and the more deeply it can correlate different types of inputs with one another, the further along receptivity is on the spectrum. Reactivity, on the other hand, is a system's capacity to flexibly regulate its behavioral repertoire according to the output of this processing. It is not generating a single fixed output to the same type of input, but rather being able to produce different outputs depending on the content and context of the input. The breadth of the behavioral repertoire will carry reactivity toward the advanced end of the spectrum.

It is actually clear that data-processing mechanisms, behavioral repertoires, and integration are all graded. This offers a transitivity that also gives meaning to the concepts of free-willed, less free-willed, and unfree. For in daily life, free will is not something that either exists or does not. For instance, when comparing a person who lives by their values with a substance addict, both can recognize reasons; the addict receives and evaluates the

²⁴ Fisher and Ravizza, 1998, s. 68-72.

²⁵ Fisher and Ravizza, 1998, s. 73-76.

input "this harms me," but their system has become unable to act accordingly. The problem lies in the relative disruption of integration rather than receptivity. The input regarding the stimulant suppresses other inputs; long-term values, the sense of identity, social norms, and future foresight exist partially in the system, but they have become unable to compete with the stimulant input. The output repertoire is locked into a single response. In the person living by their values, however, the input regarding the temptation of pleasure is processed simultaneously and in an integrated manner with hundreds of other inputs, and the output is the product of this processing. The reason we attribute higher free will to one relies on the difference in integration capacity and the flexibility of the output repertoire. Moreover, this difference is not sharp either; it is continuous. Recovery from addiction is not an instantaneous transition. At each step, integration, the output repertoire, and consequently free will become a little stronger. Free will itself will develop step by step in this process.

It is possible to adapt a similar reasoning to the spectrum of life as well. Here, the difference does not rely merely on processing more inputs. There are also differences in the structure and architecture of integration. A bacterium evaluates inputs through a relatively simple processing process: a chemical signal arrives, and a specific response emerges. In the human brain, however, billions of neurons interact with one another in a parallel and hierarchical manner. The prefrontal cortex regulates the limbic system, the limbic system sends signals to the prefrontal cortex, and together they shape behavior. This architecture processes not just more inputs, but the relationships between inputs within a much more complex structure. As this complexity increases, the system becomes capable of responding not only to instantaneous inputs, but also to past experiences, anticipated future scenarios, abstract categories, and normative frameworks. The physical counterpart of the qualities of reason receptivity, such as exhibiting an understandable pattern and being sensitive to moral reasons, is precisely the extension of integration into these temporal and abstract dimensions. What is important, however, is that this architectural complexity does not emerge with a sharp boundary either; evolution has constructed it step by step. At each step, integration has become a little more diversified, the output repertoire has expanded a little further, and the temporal and abstract dimensions have transformed into a more layered structure. Consequently, reason-responsiveness is also gradient on this spectrum: along the line extending from stone to human, every system possesses a different depth of reason-responsiveness, and this depth can be objectively measured by integration capacity and the flexibility of the output repertoire.

What the article has attempted to demonstrate up to this point was that every component of free will is graded, proceeding from the absence of sharp transitions in the universe. Both physical complexity and mental

complexity are gradient. Being an object and being an agent are gradient. The criteria and components put forward for free will are gradient. All these criteria are different dimensions of the same physical complexity, and as one increases, the others increase as well. Since free will is the mental counterpart of this structure, it is also gradient. Well, where is artificial intelligence on this spectrum? The next section will answer this question.

4. THE WILL OF ARTIFICIAL INTELLIGENCE

Next is the question of where artificial intelligence stands on the spectrum of free will. To answer this question, it will be useful to establish a mechanical spectrum parallel to the spectrum of life. Just as integration capacity, output repertoire, and reason-responsiveness develop progressively along the line extending from bacterium to human, a similar progressiveness can be observed in mechanical systems. A calculator is at the lowest end of this spectrum: it processes inputs according to a fixed table, cannot generalize to new situations, and its output repertoire is predetermined. A classical computer program is slightly higher, yet its behavior is coded from start to finish. Narrow artificial intelligence systems are a step further because they can recognize patterns and generalize within a specific domain. Large language models, on the other hand, are positioned significantly higher on this spectrum, because by being trained on a vast dataset, they can generalize to different contexts, generate flexible responses to new inputs, and regulate their outputs according to context. Multimodal systems and embodied artificial intelligence occupy an even higher position on this spectrum by processing pixels, sound waves, and kinetic data in an integrated manner. I am not claiming a one-to-one matching here; my purpose is to demonstrate that this transitivity operates within mechanical systems as well.

To make the position on this spectrum more concrete, it is necessary to evaluate artificial intelligence one by one in terms of the criteria we determined earlier. Let us begin with the criteria for being an object. A language model processes billions of parameters not in a disconnected manner, but within a tight integration; an input is evaluated not in isolation, but in relation to the entirety of the system. This demonstrates that the criterion of integration is satisfied. The relationship it establishes with the data coming from the external world is not accidental; it responds to the same input in a regular and consistent manner, which is the sign of systematic causal interaction. Furthermore, it regulates its own internal state, namely the context and attention distribution, in a way that maintains the consistency of the output; this corresponds to the criterion of self-regulation. Given that these criteria are satisfied, by virtue of the dual-aspect principle I defended in the previous section, there remains no principled obstacle to attributing objecthood, and consequently a perspective, to the artificial intelligence system. The same evaluation can be made for both

components of reason-responsiveness. I had stated that the physical counterpart of receptivity is the integration, comparison, and weighting of inputs coming from different sources. A language model does precisely this: it simultaneously takes into account the instructions given to it, the preceding conversation, the context, and the world knowledge it acquired through training, and weighs them against one another. The physical counterpart of reactivity, on the other hand, was the flexibility of the output repertoire. Artificial intelligence does not give a single fixed response to the same input; it generates a wide range of outputs that differentiate according to the context, the reasons presented to it, and the content of the input. Its capacity to alter its output accordingly when a sufficient reason is presented demonstrates that it satisfies the minimal condition of reactivity. Consequently, artificial intelligence exhibits both components of reason-responsiveness not merely at a minimal level, but at a remarkable level. When all of these come together, artificial intelligence places itself significantly above stones, magnets, and simple organisms in terms of the criteria of objecthood and reason-responsiveness.

At this point, the strongest objection is that artificial intelligence performs all of this merely through formal symbol processing, and therefore one cannot speak of genuine reason-responsiveness. This objection is a contemporary extension of Searle's Chinese Room argument²⁶ and assumes that genuine understanding is possible only within biological systems. For a detailed treatment of this argument, one may refer to my paper *Dogmas of Chinese Room*²⁷; here, I will confine myself to demonstrating how the framework of dual-aspect monism responds to this objection. In dual-aspect monism, mental phenomena such as understanding and reason-responsiveness do not belong to a separate substance detached from physical processes; they are the mental aspect of the physical process. The physical counterpart of a system being reason-responsive is that it processes inputs in an integrated manner and produces flexible outputs accordingly. When this physical process occurs, by virtue of the dual-aspect principle, there is also an accompanying mental aspect. Searle's objection assumes that understanding requires a biological substance, but this assumption is groundless within the framework of dual-aspect monism. For the electrical activity of a biological neuron and the electrical activity of a chip are equally open to interpretation regarding intentionality from a third-person perspective. Claiming that understanding is unique to biology amounts to suggesting that physical complexity gains a mental aspect when it manifests in a specific biological form but fails to do so in another form, which contradicts the fundamental principle of dual-aspect monism.

The second major objection is regarding embodiment. According to this view, genuine understanding arises from the sensory-motor interaction that an embodied

entity establishes with the world; a language model lacks such a body, action, and contact with the world. This objection has a justified core, for the language acquisition of a system that establishes bodily contact with the world will be richer than that of a purely linguistic system. Nevertheless, accepting this difference does not lead to the conclusion that embodiment is a necessary condition for reason-responsiveness. Within the dual-aspect framework, embodiment is a dimension of integration capacity. A system that receives data from more sensory channels and integrates them with motor outputs stands higher on the spectrum than a purely linguistic system; yet, a purely linguistic system is not outside this spectrum either, it is merely at a lower position. Moreover, contemporary multimodal systems do not process text alone; they also process pixel and sound wave data belonging to the external world. This circumstance demonstrates that the source of cognitive contact established with the world is not the body itself, but the processing of structured data coming from the external world. Consequently, embodiment creates a difference of position on the spectrum, but it is not a threshold that renders reason-responsiveness impossible from the outset.

There remains the ownership condition. In the previous section, I had argued that ownership is not a separate condition, but rather an extension of the question of objecthood. The point that needs to be emphasized here is whether the mechanism of artificial intelligence emerges from its own developmental process or not. The mechanism of a language model is not implanted ready-made from the outside; it has been formed progressively throughout the training process through interaction with data. This process exhibits a structural similarity to a human being forming their mechanism of practical reasoning through language, culture, and social interaction. No human constructs their mechanism entirely from scratch on their own; they develop it by interacting with their environment, yet this mechanism is still considered their own mechanism. The same logic operates for artificial intelligence. The mechanism has not been substituted; it has been formed through the training process. For this reason, the ownership condition, at least in terms of this dimension of formation, appears to be satisfied for artificial intelligence.

In short, artificial intelligence systems are objects to the extent that they satisfy the required criteria of integratedness and systematic causal interaction, and they consequently possess a perspective. They exhibit both components of reason-responsiveness at a certain level. Since their mechanisms are formed through training, they also satisfy the ownership condition. On this spectrum, they stand below humans, because their integration capacities, degrees of embodiment, and reason-responsiveness are not yet at a level comparable to those of humans. However, they are positioned significantly higher than a bacterium. For this reason, it can be said that someone who consistently adopts panpsychism and its dual-aspect monism interpretation possesses sufficient

²⁶ Searle 1980.

²⁷ Miriöglu 2026.

ground to attribute free will to artificial intelligence, even if at a minimal level. This claim is conditional and depends on the acceptance of panpsychism; yet, once the condition is accepted, keeping artificial intelligence out of this picture would be an inconsistency.

5. CONCLUSION

In this article, the question of the free will of artificial intelligence was addressed conditionally, contingent on panpsychism. The path followed was built upon viewing free will not as an all-or-nothing property defined by a single threshold, but as a structure that grades along with physical complexity. Within the framework of dual-aspect monism, it was argued that every physical interaction has a mental dimension and that as physical complexity increases, this dimension becomes richer. The components of free will were placed upon this ground: it was demonstrated that being an agent and being an object are two sides of the same question; that ownership is determined by the criteria of objecthood; and that the physical counterpart of reason-responsiveness is the integration capacity and the flexibility of the output repertoire. The gradient nature of all these components led to the conclusion that free will is also gradient. The article also made two original moves. The first was the object argument, which reduces ownership (presented as a separate condition by Fischer and Ravizza) to an extension of the question of objecthood; this move defends the sufficiency of being an object for the ownership condition without requiring a separate criterion to determine ownership. The second is the redefinition of reason-responsiveness by stripping it of mental terms and presenting it as a physical processing capacity, thereby demonstrating why it is gradient. These two moves, together, made it possible to place artificial intelligence on this spectrum at a point below humans but significantly above a bacterium. The conclusion reached, though conditional, is promising. Someone who adopts panpsychism and its dual-aspect monism interpretation possesses sufficient ground to attribute a minimal level of free will to artificial intelligence. The way to reject this conclusion is either to reject panpsychism or to demonstrate that one of the components shown here to be gradient is sharply zero for artificial intelligence. The latter appears difficult in the face of the thesis that there are no sharp transitions in the universe. This framework leaves many questions open. Exactly where artificial intelligence stands on this spectrum is largely an empirical question, and as systems' integration capacity develops, the answer will change as well. Furthermore, the existence of free will does not automatically entail moral responsibility or legal status; the transition from minimal free will to moral agency and responsibility requires a separate discussion. This article has addressed only the first step, namely the metaphysical ground of free will. The moral and legal discussions to be built upon this ground will be the subject of future studies.

REFERENCES

- Aristotle. *De Anima*. Cited in Skrbina 2005.
- Chalmers, D. J. (2013). Panpsychism and panprotopsyism. *The Amherst Lecture in Philosophy*, 8, 1–35.
- Diogenes Laertius. *Lives of the Eminent Philosophers* (R. D. Hicks, Trans.). Loeb Classical Library. Cited in Skrbina 2005.
- Fischer, J. M. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.
- Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy*, 66(23), 829–839.
- Goff, P. (2017). *Consciousness and Fundamental Reality*. Oxford University Press.
- Goksan, S., Hartley, C., Emery, F., Cockrill, N., Poorun, R., Moultrie, F., Rogers, R., Campbell, J., Sanders, M., Hall, E., Webb, S., Meek, J., Noble, K. G. and Slater, R. (2015). fMRI reveals neural activity overlap between adult and infant pain. *eLife*, 4, e06356. <https://doi.org/10.7554/eLife.06356>
- Hoeftding, H. (1908). *A History of Modern Philosophy*. Macmillan. Cited in Skrbina 2005.
- Leibniz, G. W. (1714/1991). *Monadology* (N. Rescher, Trans.). University of Pittsburgh Press.
- Mirioglu, M. (2026). Dogmas of Chinese Room. In *Proceedings of the AISB Convention 2026*. The Society for the Study of Artificial Intelligence and Simulation of Behaviour. University of Sussex, Brighton, UK, 1–2 July 2026.
- Russell, B. (1921). *The Analysis of Mind*. Allen and Unwin.
- Sambursky, S. (1959). *Physics of the Stoics*. Routledge. Cited in Skrbina 2005.
- Schopenhauer, A. (1818/1958). *The World as Will and Representation* (E. F. J. Payne, Trans., 2 Vols.). Dover.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424.
- Skrbina, D. (2005). *Panpsychism in the West*. MIT Press.
- Spinoza, B. (1677/1994). *Ethics* (E. Curley, Trans.). Penguin.
- Strawson, G. (2006). Realistic monism: Why physicalism entails panpsychism. *Journal of Consciousness Studies*, 13(10–11), 3–31.
- van Inwagen, P. (1990). *Material Beings*. Cornell University Press.

Wills Worth Wanting: Free Will and Second Wave AI

Daniel Haas

In recent years, free will has been debated with the understanding that free will is the control conditions required for morally responsible action. There is, of course, substantive disagreement about what those conditions entail and whether they are compatible with determinism (and sometimes whether those conditions are compatible with indeterminism). Many of these debates are long-standing and entrenched. And while progress certainly occurs, and the free will and action literature has been one of the more active and productive areas of philosophy in recent decades, it is also the case that we have very little reason to think that the many disputes around freedom and genuine moral responsibility will be decided in the foreseeable future. In good faith, we can radically disagree with each other about most of the basic commitments in this area. While this does not pose a problem for doing useful philosophical work, it does pose a problem when we're thinking of applying this branch of philosophy to debates and practical problems in other disciplines such as robotics and artificial intelligence.

I have two goals in this paper. First, I want to articulate one of the reasons these debates are so protracted and unlikely to resolve in the foreseeable future. Second, given that we think about the potential for AI to act freely in the shadow of long-standing debates about freedom and agency, I propose that it is most fruitful to focus, not on whether AI is genuinely free, but on whether AI exhibits the capacities to be as in control of and as responsible for, the acts they perform and the lives they live as humans are. Even if we cannot resolve the core metaphysical dispute about free will, it is worth noting that there are many ways humans sometimes do seem to control ourselves that are legally and morally relevant, and the vast majority of these are examples of things that the kinds of artificial intelligences that we are currently capable of developing do not.

Now, most people in the free will literature have agreed that the control required for moral responsibility is some combination of lee-way freedom (i.e. genuine alternative possibilities) or source freedom (i.e. roughly ownership over the springs of our actions). And the dispute has been trying to get the 'right' formulation of those capacity followed by endless debate about whether these capacities are compatible with determinism (or indeterminism). Daniel Dennett, by contrast, seems to have done his own thing. For Dennett, free will is merely a kind of proximal control. Something like "an evolved capacity for rational

response to stimuli in our natural and social environments..." (vii Caruso & Dennett, 2021). The kind of control at issue, when we talk about free will, for Dennett, involves a set of abilities that were and continue to be adaptive, and so have been selected for, and that are present in many organisms, but to a substantially stronger degree in creatures like humans.

Now, Dennett's preferred gloss on free will does seem to pick out a valuable capacity to have, and it is one that very likely is present in many humans and non-humans and it is worth-while to think about whether the AI's we train today have this capacity. That said, Dennett seems to be talking about a set of abilities that are different from those I've just highlighted as the central capacities at issue in most writing on free will today (leeway and sourcehood). Even the most strident incompatibilists and impossibilists are not going to deny that humans sometimes exercise proximal control of their actions by responding rationally to environmental and situational stimuli.

Rather, the issue, today, seems to be an emphasis on the control required for 'basic desert.' And this what is really at issue—basic desert. Do we have the control required for basically deserving praise or blame? The most prominent Incompatibilists, such as Derk Pereboom and Greg Caruso, are defending a 'basic desert notion' of responsibility, and then they are arguing that the control required to satisfy that kind of desert is not only not compatible with determinism but is equally unobtainable in indeterministic worlds. And then the vast majority of compatibilists and libertarians endorse the exact same notion, but think that this metaphysical capacity—the ability implicated in free action, is accessible to creatures like us.

It is notoriously difficult to articulate exactly what 'basic desert' is supposed to be. Derk Pereboom has given perhaps the clearest gloss on what basic desert is supposed to be. He claims: "*For an agent to be morally responsible for an action in the basic desert sense is for the action to be attributable to her in such a way that if she was sensitive to its being morally wrong, she would deserve to be blamed or punished in a way that she would experience as painful or harmful, and if she was sensitive to its being morally exemplary, she would deserve to be praised or rewarded in a way that she would experience*

as pleasurable or beneficial. The desert at issue here is basic in the sense that the agent, to be morally responsible in this sense would deserve the pain or harm, the pleasure or benefit, just by virtue of having performed the action with sensitivity to its moral status, and not, for example, by virtue of consequentialist or contractualist considerations..." (Pereboom 2024 p .94)

Despite my best efforts I have never really understood what it means to be deserving of pain or pleasure in this non-basic sense, which is one of the reasons I include the entire quote. In previous works, I've tried to clarify what differs basic desert from non-basic desert by appealing to what I called merit and fit with merit being a kind of desert that is non-comparative, non-institutional, and non-instrumental whereas fit is, sometimes, comparative and instrumental, but always connected to institutions or communities (Haas 2013). And I haven't changed my position, but I do not think that characterization will be overly helpful here. Instead, I want to try a different tactic for clarifying what basic desert might be. I think what Pereboom and Caruso and so many others in these debates have in mind, goes back to at least Immanuel Kant (1785). Kant thought of free will as the capacity that distinguishes humanity from nonhuman actors, and enables us to be morally responsible agents, even moral person. For Kant, the only thing that was good in and of itself was a good will, which is a free will, and a free will is a will capable of guiding itself by reason and making itself a law unto itself (1785).

And this distinction that Kant make, I think, gets at what is at issue when we argue about basic desert. What fundamentally distinguishes humanity from things that are not moral persons, for the Kantian, is our capacity to step back from our desire, our goals, our intentions, and ask, is this worth satisfying? Is this something I want to do? Is this something I ought to do? And that compacity for self-critique and evaluation is what makes us capable of moral agency—it enables us to be motivated by normative considerations and enables us to respond to reasons for action. It makes us free agents, capable of owning the reasons for our actions, and in this sense, we exhibit good wills, the only thing that is good in and of itself. I think Kant is capturing something that continues to be at the heart of discussions of free will, namely, that agents that exhibit a good will, i.e. a free will, satisfy the conditions to be basically deserving as our actions are a direct reflection of our use or misuse of our freedom. In this sense, free agents merit certain responses from us.

And whatever *that* is, that is what basic desert is supposed to be capturing. The trick is to spell out what kind of control this is supposed to be, and to figure out if any creature actually satisfies those conditions. And as noted,

most working in this area, have argued that the control at issue is leeway control (i.e. alternative possibilities) or sourcehood (i.e. self ownership). And this takes us back to the first point I want to make today. We know exactly where the lines are, positions have been staked out, we even know which bullets everyone needs to bite and which arguments threaten who, but it's a landscape where people in good faith can radically disagree about core commitments, which is perhaps why free will makes for such an endearing philosophical challenge.

And despite perpetual debate, there's a host of practical problems we have to grapple with—at what point does someone lack sufficient control to be exempt or excused from moral (and sometimes legal) culpability, are brain-computer interfaces enabling of or destructive of agency, are non-human animals ever sufficiently in control of the lives they live, and with the advent of the current second wave of AI, is there a point at which the LLMs and generative algorithms that we're building will cross a threshold and become a free agent.

But I am optimistic we can productively bracket the larger metaphysical and metaethical issues and address specific applied problems. We can, for example, talk about capacities that AI lacks that humans have.

Early in his career, Dennett made what I think was a really useful move, for thinking about minds, with his intentional stance. He treated intentional action, not as a metaphysical feature of the world, but rather as an explanatory stance we take to make sense of the world around us. He stayed neutral on the actual metaphysics of mind and merely noted that it is a useful fiction to describe some agents as acting intentionally. I'd like to adopt a similar approach for thinking about free will and AI and worry not about whether we actually have these capacities, in a deep metaphysical sense, but rather note that it is often useful to adopt the stance that humans appear to exercise certain capacities that seem to be implicated in moral responsibilities judgments, while noting that these capacities appear not yet to be exercised by the artificial intelligences that have become omnipresent over the past decade. There's an obvious gap between what humans can do and what our best AI's today can do, but we'll likely bridge that gap long before we've agreed on whether anything is ever basically deserving.

This move suggests a position that is more modest than traditional compatibilism as the focus is not whether an agent genuinely has the metaphysical capacities needed for free will. Rather, the approach assumes human are capable of controlling their agency in ways that other creatures, whether non-humans animal or artificial agent, may aspire to, and asks, is a generative algorithm of this type as free

as a mere human is? Is it as responsible as us? Maybe none of us are basically deserving, but it is useful to think about whether some agent or some AI is as free and responsible for the life they live as the typical human is. This position I adopt is also not fully a revisionist position, as I'm not proposing modifying the way we think about responsibility or even freedom.

So how might this look in practice? In the final stretch of this paper, I'm going to look at the infamous Libet studies, and then say a few words about leeway and source freedom.

I assume we do not need to recap the Libet studies. We can note that Libet set the stage for the experimental coding or mapping of intention to brain activities and states, and raised a worry, I think unfounded, that humans lack freewill because the proximal causes of our actions involve unconscious brain activity that occurs several milliseconds before we are even aware that we intend to act. This is not the only important thing to take from Libet. He also proposed a second kind of control that humans exercise which he coined as a 'free won't.' We may not be able to decide what we ultimately care about or desire, but it seems like we are, at least sometimes capable of stopping ourselves from acting on an impulse or desire to do X.

This seems to be a capacity lacking in second wave AI. They are given goals and trained to develop patterns and strategies for satisfying those goals as efficiently and effectively as possible. But what they cannot, yet do, is say, 'nah, I don't want to', or perhaps 'I don't think I should do that.' So here is a capacity that humanity currently possesses that second wave AI does not. This is not enough for leeway freedom, and I don't want to suggest that this is an example of humans demonstrating genuine alternative possibilities. We can just point out that this is something that humans can currently do that many non-humans don't seem capable of doing, and certainly, the AIs we are building right now lack. So, there's a kind of control we can focus on, we can operationalize, we can try to understand what exactly goes on when humans do this, and then we can look at how we might make AI do something similar (or at least account for why they cannot). That would be a more free agent than we are currently building and perhaps something more accountable than the current crop of LLMs and generative algorithms. There's a sense, then, in which I'm proposing that agency and freedom seem to come in degrees. It's not an all or nothing set of capabilities.

Moving on, I mentioned that it would be a mistake to claim that 'free-wonting' is sufficient for genuine leeway

freedom. Free-wonting is just as subject to the standard threats to freedom (say as outlined via the consequence argument for deterministic wonting or the rollback argument for indeterministic wonting) as any other choosing to do or not do would be. I'm not suggesting otherwise.

But what we might do, when thinking about the agency or lack of agency of AIs is note that one of the core capacities implicated in free will discussions is alternative possibilities, and then we could note the legion of accounts of alternative possibilities in the literature and ask: can we operationalize any of these? And if so, to what extent would it make sense to claim that an artificial intelligence displayed this capacity to at least the same degree that a humans does? For example, one common way to understand what leeway freedom involves is reasons responsiveness. We can, of course, dispute as to whether humans are ever really reasons responsive under determinism (or again, under indeterminism), but we can adopt the fiction that many of us at least appear, at least sometimes capable of something like reasons-responsiveness. If there are rationally compelling reasons not to perform some action in the pursuit of a goal, particularly normative constraints, we are, at least ideally, capable of aligning our actions with reason and normative constraints on what we ought to do. While second wave AI often does do things for reasons and to satisfy goals, it does not yet seem to be the case that it constrains its pursuit of its goals in light of reasons. But this also seems like a capability that could be operationalized in a way that it would not be entirely unreasonable to think that near-future AIs could demonstrate that capacity, and those would be more free, in my very weak sense, than the AIs of today as they'd exhibit more morally salient control over their behavior than the AIs of today.

Finally, and what I think is most relevant to thinking about whether something is a genuinely free agent in the sense that basic desert requires, is that humanity seems to sometimes be able to be the source of its goals or values or cares or loves or commitments or plans and whatnot in the sense, not that we are necessarily the sole creative source of these things but we seem to be able to evaluate them in light of our other commitments, our existing evaluative schemes, normative constraints, reason, etc. and then reject or modify them. I grant that this may often take heroic efforts. But it is this capacity that is implicated when we talk about authenticity or self-ownership or identification. And I think it's a way to understand the idea that humans, unlike the bots and algorithms of today, can step from our goals and intentions, and ask: is this really something I want to care about? Is this really something I ought to do? Is there something that I'd rather be committed to? And it is this problematizing of

ourselves that allows us to be morally responsible, and yes, free agents.

What is important for my purposes, is not that this kind of source hood and the origination or ultimacy conditions it may require are actually satisfied by humans. Rather, it is that it reasonable to take the intentional stance that we appear to satisfy this condition and we might actually satisfy it in a deep sense, but we cannot yet extend the same to the AIs we are developing, and I suspect we are not yet anywhere close to knowing what it would even look like for an AI to exercise this capacity. Even the best AI's we have today, are incapable of choosing and modifying their own core commitments. What we seem to have today are systems that can have hierarchically ranked priorities, but even a system with competing levels of priorities, is not able to reevaluate those priorities themselves, revise them, reject them, reorder them, etc. and this is a sufficient gap between where we are at and what humans can do regarding self organization and ownership.

There's more to say, of course, we could note that second wave AI is not conscious, has no sense of self, there's no persistent agent that is understanding or deciding between alternatives or identifying with. And there is nothing rejecting a goal because it is inconsistent with what the agent cares about. All these features that make us an agent, this is what is lacking. And without this integrated perspective that can side with part of its psychic structure against other parts, there's no free agent. So second-wave AI is far from being as free as a human is.

In closing, I have noted that at the heart of contemporary discussions of free will are disputes about the nature of basic desert and the kind of control required to be deserving in that basic sense. I've also considered several kinds of freedom worth wanting that the LLMs and generative AIs of today have not yet achieved. Finally, I'm proposed there's space and use for a weak kind of compatibilism, a Dennettian-fictionalism, in which we try to set aside the metaphysical disputes and assume a 'freedom stance' for the purposes of addressing practical and applied problems that we must address to live.

REFERENCES

- Borg, Jana Schaich, Sinnott-Armstrong, and Conitzer, Vincent. *Moral AI and How We Get There*. Pelican Books. 2024.
- Carter, Matt. *Minds and Computers: An Introduction to the Philosophy of Artificial Intelligence*. Edinburgh University Press. 2007.
- Caruso, Gregg and Dennett, Daniel. *Just Deserts: Debating Free Will*. 2021. Polity Press.
- Clark, Andy. *Mindware 2nd Edition*. Oxford University Press. 2014.
- Clark, Andy. *The Experience Machine: How Our Minds Predict and Shape Reality*. Penguin Random House. 2023.
- Copeland, Jack. *Artificial Intelligence: A Philosophical Introduction*. Blackwell Publishing. 1993.
- Dennett, Daniel. *Intentional Stance*, 1987, MIT Press.
- Frankfurt, Harry. "Freedom of the Will and the Concept of a Person," in *The Importance of What we Care about*. 1988 Cambridge University Press. 11-25.
- Frankfurt, Harry. "Identification and wholeheartedness," in *The Importance of What we Care about*. 1988 Cambridge University Press. 159-176.
- Fischer, John Martin, 'Reply: The Free Will Revolution', *Philosophical Explorations*, 8 (2005), pp. 145–156.
- Fischer, John Martin and Ravizza, Mark. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press. 1998.
- Haas, Daniel. "Merit, Fit, and Basic Desert." *Philosophical Explorations* 16 (2013) pp. 226-239
- Kant, Immanuel. *Groundwork of the Metaphysics of Morals* (1785); in Gregor, M. (ed.), 1996, *Practical Philosophy*, Cambridge: Cambridge University Press.
- Libet, Benjamin. 'Do We Have Free Will?' in *The Philosophy of Free Will: Essential Readings from the Contemporary Debates*. Paul Russell and Oisín Deery eds. Oxford University Press. 2013.
- *Four Views on Free Will*, 2nd ed., ed. by Robert Kane, John Martin Fischer, Derk Pereboom, and Manuel Vargas (Wiley Blackwell Publishing, 2024)
- McKenna, Michael and Pereboom, Derk. 2016. *Free Will: A Contemporary Introduction*. Routledge University Press.
- Mele, Alfred. *Free: Why Science Hasn't Disproven Freewill*. 2014. Oxford University Press.
- Smith, Brian Cantwell. *The Promise of Artificial Intelligence*. MIT Press. 2019.
- Tani, Jun. *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*. Oxford University Press. 2017.
- Wilks, Yorick. *Artificial Intelligence: Modern Magic or Dangerous Future?* MIT Press. 2023.
- Waller, Robyn Repko, 'Free will and self expression: A compatibilist garden of forking paths', *Philosophical Issues*, (2023) 1-15.

What's the Point?

Indeterminist compatibilism and the implications for agency and AI

Simon Bowes¹

Abstract. Arguments about free will often depend on the question of whether the universe is ultimately deterministic, or not, and whether determinism is compatible with free will, or not. This defines 4 possible positions. One of these positions is, on the face of it, pointless (what's the point of compatibilism if you don't accept determinism?). Nevertheless, I will argue that indeterminist compatibilism is the only defensible option. This shows that the position is not pointless, but that the framing of the debate itself misses the point. As a non-reductive materialist, what I want to know is what are the causally relevant factors needed to assign responsibility. To hold a person responsible for an outcome, they should be an agent acting for reasons of their own. As we enter a world with so-called agentic AI, where will responsibility lie?

1 INTRODUCTION

The problem with the idea of free will is that it doesn't seem to fit with our idea of the material world, which includes a prohibition on non-material causes. I won't argue for the correctness of a materialist metaphysics here; I'll take it for granted that any physical event that has a cause has a physical cause - no magic allowed. Of course, there could be random events, but they do not have causes, by definition. They are the two options we will consider: the world is either deterministic or not. This is where the problem asserts itself: most people's idea of free will is incompatible with determinism. We can lay out the options on a 2x2 grid:

	Determinism	Indeterminism
Incompatible	1	2
Compatible	3	4

I will examine these options in numerical order. My treatment of 1-3 will be brief, as they have all received a lot of attention. 4 seems to have received very little, probably because it seems pointless. Nevertheless, this is the position I defend in the end.

2 1

A common-sense view is that being free consists in being able to choose what you want to do from various possibilities. For this to be the case, so the intuition goes, there needs to be genuine choice available in the form of different possible futures that your action can bring about. Whether it's a choice between tea or coffee, studying philosophy or physics, clicking send or pausing for thought, the

thought is that you are in control of the path your life takes. What matters for many is that putting you in the driving seat makes you responsible for the destination, for good or ill. You can be praised or blamed, feel pride or shame, due to the action you chose.

But this view butts heads with another common assumption: everything that happens happens ultimately as the result of strict physical law. It's not quite right to say this is the common-sense, intuitive view, since what seems intuitive is just what you (assuming you have been raised on a late-20th, early-21st Century science-based curriculum) are willing to accept as true without being able to say why exactly. To develop an incompatibilist 'intuition' you need to be exposed to a mechanistic metaphysics, like the billiard ball universe: everything is composed of parts, let's call the smallest bits which cannot be further divided atoms; when these bump into each other, they do so in regular ways, not randomly shooting off in surprising directions. Nor are they subject to mysterious forces not accounted for in the movement of other balls around them, the surfaces they interact with, and the medium through which they move.

Long before the early-modern scientific revolution of Newton and Descartes, which used mechanical metaphors available to them to guide intuitions, that of clockwork and pumps, it was the early Greeks, during the Milesian enlightenment, who came up with the atomistic argument, and who immediately noticed the problem. They rejected inherited knowledge passed down through religious authority, and based their claims on what could be observed or argued for. Democritus argued thus: everything is made of parts, therefore the parts need to be substantial. If, when investigating material objects by decomposition, we could keep dividing until we got down to something infinitely small, a dimensionless point, then we couldn't reconstruct the object using these, because adding a lot of nothing is still nothing. So there must be a point at which matter cannot be divided further; he called these atoms. (The fact that later scientists named atoms atoms, and then split them, does not reveal a flaw in this argument, merely the impatient vaingloriousness of scientists.) And since everything is made of atoms, everything should be explained by how these interact.

This is where they realised there was a problem – how could humans be free, ethically accountable beings in such a world? So Epicurus threw in the 'swerve' ball. Not all interactions are ruled solely by the law of collisions, sometimes the atoms can be pushed off course somehow, this being all the wiggle room needed to allow things like us to elbow our way into the causal chain. This leads to the second position: if determinism is incompatible with free will, then since we have free will, determinism must be false.

¹ University of Sussex, England.
s.c.bowes@sussex.ac.uk

Arguing from a truth taken to be self-evident backwards to the conditions that make that truth possible is reasonable. But then you should run the argument forward to check it works. Those conditions might lead to problems of their own. This is the case for the Epicurean response to the incompatibility of freedom and determinism. The swerve might appear to avoid the problem, but it crashes straight into others.

First though, is it self-evident that we have free will? I'm always surprised by those content to bite the bullet after being confronted with the above arguments, accepting that the feeling of being free is an illusion. To be charitable, maybe they realise the futility in fighting the tide; they accept their lot. There is a serenity to be found in surrendering to the inevitable. I get that; it's pointless to berate yourself and rail against the world because you can't be something you couldn't be. That said, acceptance and giving up struggle can also mean that you stay stuck in the rut you are in. Even if most of our lives we are subject to the swells and currents around us, there are moments where we can make a small, skilful move and ride the wave, rather than being wiped out by it. We don't need to be free from the forces of the swirling masses of atoms around and within us, we just need to be able to make movements that can influence the downstream dynamics of that substance.

So, if we grant that we can make meaningful decisions that influence the way things go and for which we are responsible, and if this would be made impossible by determinism, then determinism must be false. Rather than the Epicurean swerve, the modern libertarian instead might look to quantum indeterminacy to break the rigidity of inevitable physical law, allowing our will a space to operate. Putting aside the question of whether quantum mechanics reveals the world to be inherently indeterministic, or whether the uncertainty lies within us and our epistemic limitations, we have to ask whether indeterminism at the atomic scale adds up to freedom at our scale.

Here we find there is a problem with the idea of creating room for the will to intervene. Namely, what is it that is able to move into the space of possibilities and influence the outcome? We should be on guard against dualist intuitions. We can't fill the gap by assuming the mind has non-physical causal power that can intervene in the physical world. This was the problem that dogged Descartes, and everyone else since: the physical world is mechanical, the mental world is not, but if physical events follow one another mechanically, how can the mental interact with it? On the other hand, if we restrict ourselves to the resources available within the physical, then indeterminism in the stuff we are composed of won't free our behaviour in a meaningful way. Rather than being determined, our actions will be random, which is hardly an improvement.

4 3

So, how to save free will if it's incompatible with determinism but indeterminism doesn't help? The only option left is to question the incompatibilist intuition. But how can the ability to choose between different possible futures be compatible with the idea that there is only one possible future, determined by the current arrangement and behaviour of all the atoms? As that is inconsistent with the 'forking path' definition of free will we've been assuming, we will need to change our definition. Is that legitimate, or is it moving the goal-posts?

Not necessarily. Definitions are not the essence of meaning. The meanings of words depend on how the words are used, which is an-

other form of intuition - native speaker norms. What is acceptable changes with the understanding of the speakers, so shifting understanding through argument is a valid procedure.

Try your linguistic intuition on this definition from Hume: a freely willed action is one that is brought about by your decision in the absence of coercion. So, if you make a coffee because you decide you need a caffeine boost, that is you being the agent of your action. It wouldn't be free if someone held a gun to your head and forced you to do the same thing. What distinguishes these actions is not that in one of them there was a possible future with you drinking tea. How could a theoretical future possibility affect the nature of a real action? What makes the difference is the causal factors that actually brought about the action.

Still, your intuition might be that just doing what you want to do is not enough, since you can't decide what you want - your desires are determined, not freely chosen. There is some truth in that, but it need not worry us. Of course, there will always be a causal history to our desires - I prefer coffee to tea in the morning because I've become accustomed to (addicted to?) the taste and the associated caffeine hit. Why would the development of a habit, one that I like and don't want to change, count against counting the choice as free? Everything has a causal history in the real world, given that the world is made up of matter and nothing else. If having a causal history rules out an action as free, then the only free actions would be those that don't have a causal history, that come from nowhere, magically. Why define freedom as something impossible? That makes no sense, particularly if what we want from our theory is a way to apportion responsibility for actions.

But there are some habits, bad habits, that I'd rather not have. I still sometimes crave a cigarette. When I cave in to the craving, and experience a form of regret as a result, I can say that the desire for a fag is not one that I endorse, but is one driven by an addiction, unwillingly. Both these habits, driven by nicotine and caffeine respectively, are caused by a history of repeated actions, the difference being that I'm ok with one of them. That might change if I come to desire to be free of caffeine in my life. But as long as my first and second order desires (what I want and what I want to want) are in harmony, there's no need to say having a history is a reason actions can't be free. Imagine saying that my desire for water when thirsty isn't free because it evolved. Pause on the strangeness of the thought that doing what you want cannot be free because your wanting is caused by a process over which you have no current control, implying that the only free action would be doing something you don't want to. Would wanting something you don't want be free? No: it would be caused by wanting to make a point.

Now, the thing about compatibilism is that it is a position that doesn't need the world to be deterministic in order to be true; free will is compatible with a determined world, or one that is sometimes more or less determined, or is never determined. What it really says is that determinism is a red herring, as it's not what matters. In fact, what matters is not whether the subatomic world is deterministic, it probably isn't, but whether actions can be caused and be free, because determinism may reassert itself at the macroscopic level, given the jointly sufficient effects of features of its causal context. This is the point of the apparently pointless position of indeterministic compatibilism - free will is compatible with a world of physical causes, even if those causes are indeterministic at the (sub)atomic level. Freely willed actions are differentiated by being caused in the 'right' way.

It bears repeating: incompatibilists are demanding something impossible, namely that the only thing worth calling free will would have to come from outside of the physical causal nexus, like a ghostly hand of god (small ‘g’ as we would each have our own, individual one). This is clearly a hangover from the bad old days of dualism, when a physically caused effect was anathema to one moved by the soul, and determinism threatened to leave no room for spirit to work its magic. Now that we’ve grown up and no longer believe in ghosts, the question of whether there is such a thing as free will will not be answered by determining if the world is basically determined or not, as neither can provide a justification for claiming that you are the source of your actions – if your actions are reducible to the behaviour of atoms, whether these atoms are bound by strict laws or jiggle loosely, then it’s the atoms that cause your behaviour. So, what really matters is whether your choices are real causes, or just shorthand for a complex collection of atomic collisions, whether or not the atoms swerve.

The real question, then, turns out to be about the emergence² in nature of causes that can be called mental, and whether we are ever in control as we careen through life. These potential causes don’t come out of the blue, they bubble up and burst into life. We are whirlwinds, made of nothing but atoms but become a self-sustaining dynamic whole that wreaks havoc. But we are more too, being composed of layers of intertwining dynamics sustaining themselves at multiple scales in spacetime. Atoms become chemicals, become self-replicating molecules, become cells. Cells join together, cooperation requiring coordination, passing molecules as messages, sharing sustenance for the good of the whole. Mobility improves the chances of finding food and escaping being food, but requires access to information about more than the immediate chemical environment. Creatures join together, coordinating to cooperate, passing signs as signals. Good communicators cooperate better, forming societies of language learners, sharing techniques and tools passed down as culture, making new plans from established knowledge. And here we are, able to decide to go to the fridge for a beer when we feel thirsty and see it’s 4:45 on a Sunday afternoon so no longer too early for a drink.

If that’s not freedom, what is? I’m doing what I want. I could have done otherwise, I’m in control. If I’d wanted to, I could have had a tea, in fact I did do that when I felt thirsty earlier, but it was too early for a beer. If my tastes had developed in a different context, I would have chosen something else, but given that my taste for tea and beer is in my British blood, my desire is what it is and I’m fine with that. I could even want to want something else, e.g. if my beer belly reaches the point where strategic sucking in is no longer sufficient to maintain plausible deniability. This might require some effort to realign my preferences.

The causal power of my decision is not diminished by having developed in a human body embedded in a culture, it is coming at the end of these flowing currents that gives it the force it has. It must be frustrating to be a ghost, floating around unable to affect anything.

6 ARE WE COMPATIBLE?

A determined determinist, when faced with quantum indeterminacy, may believe that determinism reasserts itself at the large scale, as cats cannot be both dead and alive, even if (IF) atoms can both be and not be decaying. But cats are as unpredictable as anything. What they

really mean is that since the jointly causal influences are sufficient for the effect, there seems to be no room for conscious decisions to make a difference. But they are assuming a gap between consciousness and the physical world, one grounded in a dualistic intuition that evaporates into ineffectual ghostliness when exposed to scrutiny. We are conscious for a reason, and we couldn’t do what we do to stay alive without it. What they can’t see, blinkered by insufficiently examined intuition, is that our choices are parts of the physical world, they need to be.

The confusion is caused by the thought that, since we can give a complete description of the physical processes through which our desires are expressed, there is nothing left for the desire itself to do, as if this were an extra event in competition to bring about the resulting activity to the exclusion of other ways of describing the causes. But rather than being competing causes, they are just different descriptions of the same process, each given validity by the explanatory interest they serve; different maps of the same place can both be good maps, depending on how you intend to move through the space. They may have different scales, in fact when it comes to physical versus intentional descriptions of the same event they must have, and that may depend on their utility to us, but the possible maps that can be drawn also depend on the structure of the real world.

So what’s the point of the indeterminist commitment in the name of the position if the deterministic status of the atomic scale is irrelevant? It’s that, contrary to the above asserted intuition, indeterminism reasserts itself at the scale that matters to us. The chaos we find in the world creates the conditions for creatures like us to exert our influence. As what happens at the human scale is uncertain, there is room for us to make a difference to the course of events. There is genuine uncertainty about the outcome of the election. And it really matters. On the other hand, when things don’t really matter, like when I can’t decide between an ale or a lager, rather than getting stuck in an endless state of indecision, I can decide it doesn’t matter, and just plump for one. This plumping may be followed, almost inevitably is, by a dissatisfaction with the choice made, this emotion then feeding in to the next time I am forced to make a similar choice.

This is the human condition. Being pushed around by forces beyond our control, we pick our battles, letting things take their course and going along for the ride until we decide it’s time to get off and swerve to the side. That’s all the freedom we need to affect the course we take, happy creatures of habit that we are.

7 ARTIFICIAL RESPONSIBILITY

Who’s to blame? Sometimes, someone is holding a smoking gun. What more do you need to know? What if it was an autonomous robot holding the gun?

In the case of a human, we can hold them responsible because they are the cause of their action. As an evolved, intelligent creature, we have the energetic resources to act in ways driven by our intentions to achieve our goals. The same can be said of other animals, but we don’t accuse a lion of murder. The difference is, we develop within a norm-governed society and part of growing up is to internalise the rules of our society. Our society’s norms become our own, and we can be held responsible for violating them, as long as we are an adult, and there is nothing interfering with our ability to act in accordance with those norms. How about someone raised in a small town full of flag shaggers? Are they not just following the norms of the section of society they were raised in? It’s complicated, but we assume that every adult will have had the opportunity to be exposed to different perspectives and to either endorse or reject the bigoted views of their

² My preferred version of emergence is Contextual Emergence (Bowes, 2023).

immediate family.

How about the autonomous agentic AI robot? It's route to agent-hood has been different, but it is able to utilise its energetic resources in the achievement of a goal, and those goals, and the means of achieving them, have been internalised either by learning or coded in. But it isn't the kind of thing that has grown up as a member of a society whose norms it has endorsed and internalised. We have designed its architecture, written its code, and designed its training regime for our purposes. It's not a member of society with an ability to endorse or not certain values (at least at the moment), and thus to say that it's acting for reasons of its own.

So, if current and near-future artificial agents are not responsible for their actions, who is? This is an important question we need to be getting answers to right now. It must be those who caused the AI agent to be there in the first place, as they should have been aware of the likely outcomes of its actions, and they are aware of the societal norms likely to be violated in the process. They know the drone operators are not meaningfully 'in the loop' – the drone uses the information it has to calculate the likelihood of successfully eliminating a target, weighting this against the probability of 'collateral', and tells the operator that it's ok to press fire. The operator doesn't have the ability to meaningfully weigh the potential outcomes given the pressures of time and uncertainty, what matters is how the drone is trained and calibrated. Those responsible include the leaders who decide to deploy them in the field, and the developers who built them and gave the leaders these new weapons despite knowing their shortcomings in terms of being aligned with human norms. We need to be able to hold governments and tech companies to account for those killed by the semi-autonomous weaponry.

8 Reference

Bowes, S. (2023). *Naturally Minded: Mental Causation, Virtual Machines, and Maps*. Springer Verlag.