Can AI systems have free will?

Christian List*

First version: November 2024 / this version: March 2025

Abstract: While there has been much discussion of whether AI systems could function as moral agents or acquire sentience, there has been very little discussion of whether AI systems could have free will. I sketch a framework for thinking about this question, inspired by Daniel Dennett's work. I argue that, to determine whether an AI system has free will, we should not look for some mysterious property, expect its underlying algorithms to be indeterministic, or ask whether the system is unpredictable. Rather, we should simply ask whether we have good explanatory reasons to view the system as an intentional agent, with the capacity for choice between alternative possibilities and control over the resulting actions. If the answer is "yes", then the system counts as having free will in a pragmatic and diagnostically useful sense.

1. Introduction

The increasing use of AI systems in society raises many questions: Are such systems safe? Do they behave ethically? Should they be given a legal status of their own, and who is responsible when they cause harms? Could AI systems even become conscious?[1] But despite the explosion of work on these and related questions, one question has received surprisingly little attention: can AI systems have free will? Free will is associated with autonomous agency and is often considered a precondition for moral responsibility. The question of free will thus matters for debates about whether AI systems could ever count as responsible agents in their own right.

In this article, I will discuss what it would take for an AI system to have free will. To determine whether a system has free will, I will argue, we should assess the system by reference to a checklist of three conditions: "intentional agency", "alternative possibilities", and "causal control" (List 2019). If a system meets all three conditions, then it qualifies as having free will in a practically useful sense. On this basis, I will suggest that free will in AI is much less far-fetched than perhaps expected. My analysis is inspired by Daniel Dennett's work (especially his work on the "intentional stance" but also his work on free will, e.g., Dennett 1984, 1987, 2003), though my analysis goes beyond Dennett's in a number of respects.

Among the few prior works discussing free will in AI, most employ more demanding conditions for free will and typically reach more negative conclusions (see, e.g., Floridi and Sanders 2004, Krausová and Hazan 2013, Farnsworth 2017, Sanchis 2018, and Blum and Blum 2022).[2] Others focus on the public's *beliefs* about whether AI systems have free will (Astobiza

* I am very grateful to Levin Hornischer, Sven Nyholm, Silvia Milano for very helpful comments and discussion.

[1] For discussion, see, among many others, Chalmers (2023), Delcker (2018), Dubber, Pasquale, and Das (2020), Floridi and Sanders (2004), Floridi (2023), Fossa (2018), Matthias (2004), Nyholm (2020), and Solum (1992).

[2] Floridi and Sanders (2004) seem to require special human-like internal states for free will and argue that "there is substantial and important scope for the concept of moral agent not necessarily exhibiting free will or mental states" (p. 351). While Farnsworth (2017) also emphasizes the importance of choice-making for free will, he suggests that "[t]he main impediment to free-will in present-day artificial robots, is their lack of being a Kantian whole" (p. 1). Both Krausová and Hazan (2013) and Sanchis (2018) relate free will to unpredictability. As

2024), which also suggest a broadly negative answer. The closest precursor to the present analysis can be found in Maier (2023), who, like me, argues for the possibility of free will in AI by pointing out that AI systems can be viewed as choice-making agents to which decision-theoretic models are applicable. I will develop this idea in terms of the above-mentioned checklist of three conditions, whose application to AI was also briefly anticipated in List (2019). My analysis is consistent with Dennett's idea that we should think of free will, not as requiring any unrealistically demanding powers, but as an evolved and adaptive capacity that is part and parcel of the kind of agency that we find in humans and other complex animals.

2. What is AI?

Artificial intelligence can be defined as the capacity of an artificial system, such as a computational or robotic one, to perform cognitive tasks and/or to interact with the environment in ways traditionally associated with human or animal intelligence. In line with this definition, the academic field of AI has been characterized as "the study of agents that receive percepts from the environment and perform actions" (Russell and Norvig 2021, p. 7).

One common distinction is that between "weak" and "strong" AI. "Weak AI" refers to artificial intelligence narrower than human intelligence, for instance due to being restricted to fixed computational tasks, while "strong AI" refers to artificial intelligence more similar to human intelligence in its flexibility or generality. "Artificial general intelligence", moreover, refers to a form of artificial intelligence on a par with or stronger than human intelligence across many tasks. Examples of weak AI systems include chess-playing computers and smart route planners. Strong AI is increasingly becoming a reality, as exemplified by generative AI chatbots that can conduct complex conversations or compose texts on many subjects. The quest for artificial general intelligence is the industry's next frontier, although there is no consensus on how close we are to achieving it and how desirable it would be.[3] But even current AI systems make or participate in decisions that used to be the exclusive domain of humans, whether it is driving decisions in transportation, diagnostic decisions in medicine, financial decisions in banking and investment, juridical decisions in legal contexts, and targeting decisions in the military.

---

explained in section 5.2 below, I do not agree that unpredictability is required (or even helpful) for free will. Blum and Blum (2022) define free will as "the ability to violate physics" (p. 2) and take the view that "[a]ll theory is against the freedom of the will" (p. 3). However, by appealing to the concept of a "Conscious Turing Machine (CTM)", they argue that "[a]lthough a deterministic entity living in a deterministic world cannot have free will as it is typically understood, it is entirely possible for such an entity to rightly and firmly *believe* in its own free will" (p. 1, emphasis added). More congenially from my perspective, they define "the exercise of free will" as "the act of choosing between two or more options" (p. 2) (adding "whichever option is believed to have the greatest utility") and suggest that "[t]his is something that any animal or machine with a CTM brain can do" (ibid.).

[3] While Luciano Floridi (2023) has suggested that current AI systems are best viewed as displaying a new form of agency without any human-like intelligence, Blaise Agüera y Arcas and Peter Norvig (2024) have argued that "decades from now, [today's most advanced AI large language models, including ChapGPT] will be recognized as the first true examples of artificial general intelligence": those systems flexibly cover many topics, perform many tasks, across different languages, can process not just images and text, but also audio and video, are connectable to robotic devices, and have advanced learning capacities. For discussion, see also Ananthaswamy (2024).

It is tempting to characterize AI by reference to the underlying technology. "Symbolic AI", the dominant approach in the second half of the twentieth century, refers to the implementation of AI through the explicit processing of symbolic representations, using tools from logic. "Subsymbolic AI", which is now dominant, refers to the implementation of AI through some lower-level architecture, such as a neural network, which does not by itself come with a symbolic interpretation. The leading version of this approach, "generative AI", implements AI by means of machine-learning algorithms that can generate new content, prompted by certain inputs, on the basis of statistical patterns picked up from training data. However, while the technology is important, I suggest that the *definition* of AI should focus on the cognitive and agentive capacities achieved rather than on the technology used to achieve it.

3. What is free will?

Free will can be defined as "a particular sort of capacity of rational agents to choose a course of action from among various alternatives" (O'Connor 2010; for an overview, see Kane 2002). According to our commonsense understanding, we human beings have this capacity. It is your own free choice, for instance, to read this article or to refrain fom reading it. If you have started reading it, it is up to you whether to continue or stop.

What does free will require? Free will is sometimes characterized, especially by sceptics, in ways that make it seem mysterious (see, e.g., Harris 2012, Sapolsky 2023). For instance, if free will is taken to require the ability to make "contra-causal" choices – choices that are unconstrained by the laws of nature or that even overrule them – then free will seems in conflict with a scientific worldview. Similarly, if free will is taken to require control, not just over one's actions but also over their entire causal pre-history, including everything that has shaped one's personality, preferences, and beliefs, then free will also seems impossible.

In real life, however, we are not normally interested in any such unrealistically strong capacities. The "varieties of free will worth wanting", as Dennett (1984) calls them, don't involve such capacities. When we ask in everyday contexts whether people choose and control their actions or when we assess their responsibility for their actions, we simply refer to the kinds of ordinary agential competences that humans tend to acquire in their pathway towards becoming competent adults. When a judge takes someone's free will to be a precondition for legal responsibility – for anything ranging from a breach of contract to criminal conduct – the judge does not require that the person can break the laws of nature or had control over their entire character-forming process (see Moore 2020). That would rule out legal responsibility from the outset and rob the idea of free will of its practical usefulness. Rather, the judge is concerned with the ordinary exercise of choice and control that we attribute to each other in commonsense psychology.

We distinguish between a premeditated crime, committed by someone in full possession of their cognitive and agentive capacities, and an accidental harm caused by a sleep-walker. In the premeditated case, judges attribute the action to the person's free will; in the sleep-walking case, they don't. Similarly, we distinguish between a competent adult whose ability to act is not physiologically or psychologically compromised and someone who is intoxicated or acts

out of compulsion. The former is held responsible for their action, the latter not. A useful understanding of free will – the kind of "free will worth wanting", as Dennett calls it – should support those distinctions and not rule out free will from the outset. Free will, Dennett (2003) suggests, is a biologically evolved capacity that is part and parcel of human agency. It is among the features that enable humans to navigate complex environments in a flexible and adaptive way.

Consistently with these ideas, I find it helpful to define free will in terms of three conditions (List 2019):

Intentional agency: Any bearer of free will is an intentional agent, i.e., an entity capable of acting in a goal-directed manner, based on intentional states such as beliefs and desires.

Alternative possibilities: Any bearer of free will sometimes has alternative possibilities to choose from, i.e., different courses of action this entity could take.

Causal control: Any bearer of free will has relevant control over the actions taken, in the sense that the entity's intentional states are the difference-making causes of those actions.

I will assume that these three conditions – perhaps with some further fine-tuning – are jointly necessary and sufficient for free will. That is: whenever someone or something qualifies as an intentional agent, has the ability to choose between alternative possibilities, and has sufficient control over the resulting actions, he, she, or it counts as having free will. If one or more of these conditions are violated, then not. The distinctions drawn by judges illustrate the present conditions. Someone in full possession of their cognitive and agentive capacities normally meets them; someone sleep-walking does not. A competent adult signing a contract in normal circumstances presumably meets all three conditions; someone who acts out of compulsion or while intoxicated does not.

Some philosophers, including Dennett in some of his writings (e.g., 1984, ch. 6), suggest that alternative possibilities are not needed for free will (see also Frankfurt 1969 and the overview in Kane 2002). The idea is that free will is compatible with the lack of alternative possibilities as long as the agent genuinely intends their action and appropriately qualifies as its "author". Dennett (1984, ch. 6) gives the example of Martin Luther, the church reformer in the 16th century, who, when asked to either reaffirm or renounce his criticism of the Roman Catholic Church, reaffirmed that criticism and allegedly said "Here I stand; I can do no other". Dennett notes that we regard this as a free choice on Luther's part, despite his apparent inability to act otherwise. After all, Luther "owned" his action, even if he could not have acted otherwise. However, I think that Luther should not be interpreted as having been literally incapable of acting otherwise. Rather, he could not have acted otherwise without abandoning his values, and he was not prepared to do that. He had a genuine choice, albeit one in which he strongly endorsed one of the options and equally strongly rejected the other (for discussions of this example on which I draw, see also Kane 2002, ch. 1, and List 2014).

For this reason, I side with those who think that free will requires alternative possibilities. This position is held not only by so-called "libertarians" (such as Kane), who think that there could

not be any free will without a form of indeterminism or openness of the future, but also by those "compatibilists" who define the notion of "having alternative courses of action" in a way that is compatible with determinism.[4] Dennett, in other writings, suggests that there is a sense in which determinism does not imply "inevitability" (Dennett 2003), and that there is "broad" or permissive interpretation of "possibility" relative to which an agent could be said to have alternative possibilities even in a deterministic world (Taylor and Dennett 2002). (More on this in section 5.2 below.) Indeed, common sense tends to represent free choices as requiring a "fork in the road", where the agent could do one thing or another (e.g., Carstensen et al. 2023).

Free-will sceptics claim that people never have all three capacities required for free will: intentional agency, alternative possibilities to choose from, and causal control over their actions. What we conventionally consider a free choice, the sceptics say, is no more under our control than bodily reflexes or compulsions. For the sceptics, free will is an illusion: everything is the consequence of a physical system inexorably evolving under the laws of nature (Pereboom 2001, Sapolsky 2023).

This article is not the place to respond to such free-will scepticism in general. Recall, however, that free will, when understood in the present down-to-earth sense as the capacity for choice-making agency, is plausibly something that has biologically evolved under the pressure to navigate complex environments (Dennett 2003). Furthermore, society, including our legal system, tends to assume that human beings normally have the three capacities required for free will: agency, choice between alternative possibilities, and control over their actions. The sciences of human behaviour operate on this assumption, too (List 2019, 2023a). Disciplines ranging from economics and political science to anthropology and history all take what Dennett (1987) calls "an intentional stance" towards the people they study, assuming that *they are intentional agents who make choices and have a relevant form of causal control over those choices*. Those sciences thereby presuppose that people have free will under the present definition. Call this the "free-will presupposition".

If we wanted to eliminate this presupposition from our explanations of human behaviour, we would need to change our entire explanatory paradigm: *all* references to agency and choice would have to be replaced by references to external factors and impersonal causal processes. We would have to abandon the "intentional stance" in favour of a "physical stance". We would no longer be talking about agents making choices. Rather, we would have to reconceptualize people as mere physical systems or passive spectators: organisms that are moved by factors beyond their control, just as air molecules float around in a thermodynamic system. *Intentional explanations*, which depict people as goal-directed agents who make intelligible choices between different possible actions, would have to be replaced by *dynamic or stochastic explanations*, along the lines of how we explain the motion of the planets, heat diffusion, or fluid dynamics. Arguably, the fact that intentional explanations are so central to so many explanatory practices – from commonsense psychology and our legal system to the human and social sciences – lends further support to free will as a working hypothesis (List 2019, 2023a).

[4] For a discussion of the available strategies, see List (2014).

4. Free will in AI

It may seem far-fetched to look for free will in a system that is, at bottom, nothing more than a digital computer interacting with its environment. Indeed, if free will required some mysterious property such as an inner "homunculus" or the ability to transcend the laws of nature, then AI systems would be unlikely candidates for having free will. However, so would humans. As Dennett (2003, pp. 2-3) notes,

> "What you are is an assemblage of roughly a hundred trillion cells, of thousands of different sorts. … Each of your host cells is a mindless mechanism, a largely autonomous micro-robot. … The more we learn about how we have evolved, and how our brains work, the more certain we are becoming that there is no … extra ingredient. We are each made of mindless robots and nothing else, no non-physical, non-robotic ingredients at all."

So, if we are prepared to say that human beings – as composite systems consisting of seemingly mindless building blocks – have free will, then we must also accept that the underlying mechanical make-up of an AI system does not by itself rule out that such a system has free will. On the analysis proposed here, to determine whether an AI system has free will, we must use the checklist of the three above-stated conditions to assess the system: an AI system has free will *if and only if* it has intentional agency, alternative possibilities to choose from, and causal control over its actions (List 2019).

But how do we establish whether a given system meets those conditions? Here, Dennett's methodology from his work on the "intentional stance" is helpful. To determine whether a given system is an intentional agent, Dennett suggests, we should not get too carried away with deep metaphysical questions but ask whether the system can be *well-explained by viewing it as an intentional agent*. For Dennett (2009, p. 339), "[a]nything that is usefully and voluminously predictable from the intentional stance is, by definition, an *intentional system*". On this account, to be an intentional agent is to be a system that can be *well-explained* by taking an intentional stance towards it. Similarly, one might say, anything that is well-explained by viewing it as a system with intentional agency, alternative possibilities, and causal control over the resulting actions should count as having free will.

Dennett's account may seem to suggest that whether a system *is* of a certain kind boils down to whether the system is *interpretable* as being of that kind. This seems to represent the system's capacities – such as agency and free will – as being largely in the eye of the beholder. Many people will find such an account too interpretivist. However, Dennett's intentional-stance criterion can also be combined with a more realist account of a system's capacities, according to which these capacities are real phenomena and not just in the eye of the beholder. Such an account would treat the availability of a certain kind of explanation of a system as an *indicator* of the system's capacities, even if it is not the *defining criterion* for those capacities.[5]

[5] Note that an *indicator* for something need not be the *defining criterion* for it. For instance, my possession of a driving licence is an indicator of my ability to drive, but it is not the defining criterion of that ability.

For example, if a system can be well-explained from the intentional stance, this is evidence for the hypothesis that the system is really an agent. (Compare the following: the fact that physical systems can be well-explained by assuming that there is gravity and electromagnetism is evidence for the hypothesis that there really is gravity and electromagnetism.)

In line with this, I propose that, to determine whether a given system has free will, we should ask whether we have *good explanatory reasons* to view the system as an entity that meets the three above-mentioned conditions, i.e., as (i) an intentional agent with (ii) alternative possibilities to choose from and (iii) control over its actions. I will now run through these conditions and indicate what would be required for a positive answer.

4.1. Intentional agency

An *intentional agent* is an entity capable of acting in a goal-directed manner, based on intentional states such as beliefs and desires. Following a long-standing tradition in philosophy and computer science, we can define *beliefs* as representations of what things are like and *desires* as representations of a target state of things to be achieved or as rankings of different such states (or assignments of utility to them).[6] Given these definitions, we have good reasons to view many AI systems as agents with beliefs and desires. If we didn't take an intentional stance towards those systems and focused solely on their low-level algorithms, we would not adequately capture their cognitive and agential capacities. Recall that Russell and Norvig (2021) characterize AI in terms of agency. AI systems, they remark, "operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change, and create and pursue goals" (p. 4).[7] Indeed, the agentive and cognitive capacities of AI systems have advanced dramatically in recent years, and more and more such systems seem to warrant an agential description.

That said, there is a lively debate on whether attributions of beliefs and other intentional states to AI systems are genuinely justified. It may be objected, in particular, that the surface-level outputs of AI systems cannot generally be interpreted as accurate reflections of stable internal states that play a belief role. For example, a large language model may give inconsistent responses to different users or in response to different prompts, and it may merely function as a "stochastic parrot" (Bender et al. 2021) whose outputs statistically fit its inputs, relative to some loss function. This would speak against the view that the system genuinely has the kinds of belief-and-desire states that are required for intentional agency.

However, as Levinstein and Herrmann (2024) have argued, "our best theories of belief and decision making make it a very live possibility that LLMs *do* have beliefs, since beliefs might very well be helpful for making good predictions about tokens" (p. 5). Even if the underlying algorithm simply leads the system to produce outputs that minimize some loss function, without any reference to anything like belief, truth, consistency, or representation, it could still

[6] On belief-desire agency, see, e.g., Bratman (1987).

[7] The intentional stance is also evident in Russell and Norvig's more detailed taxonomy of the capacities of AI systems, where they distinguish between *thinking* humanly and/or rationally and *acting* humanly and/or rationally.

be that, as a byproduct of this minimization exercise, some internal states of the system come to play the role of representations or beliefs. Levinstein and Herrmann note: "[i]t is easy to generate decision contexts (such as strategic board games, investing, figuring out how to get to Toronto from Prague, etc.) that do seem to push us [humans] to form accurate beliefs about the world" (p. 7), and similarly, for many problem-solving tasks, AI systems would benefit from having internal states that play a belief role. So, for an AI system, like for humans, "it is very useful to have an accurate map of the world, in order to guide action" (ibid., p. 8). Levinstein and Herrmann take this to suggest that it is "largely an empirical matter" whether AI systems have beliefs and perhaps other stable intentional states (ibid.).

Some AI researchers have further suggested that AI systems develop "world models", i.e., the sorts of representations of the environment often associated with intentional agency (cf. Ananthaswamy 2024). Gurnee and Tegmark (2023), for instance, studied LLM systems trained on spatial and temporal datasets (containing data about the world, the United States, New York City, historical figures, news headlines etc.) and concluded that the systems learned representations of space and time that were "robust to prompting variations and unified across different entity types (e.g. cities and landmarks)" (p. 1). They further found "individual 'space neurons' and 'time neurons' that reliably encode spatial and temporal coordinates", and cautiously concluded that "modern LLMs learn rich spatiotemporal representations of the real world and possess basic ingredients of a world model" (ibid.).

Now a critic might still say: perhaps it is a *useful heuristic* to talk *as if* AI systems were intentional agents, but in reality they are just mechanistic devices. Their algorithms leave no room for real intentional agency. Dennett would presumably reject this criticism, because, on his account, if ascribing intentional agency to a system is sufficiently useful for explaining the system's behaviour, this settles the question of whether the system is an intentional agent.

But we can say more. The present criticism fails to acknowledge that there are different levels at which we might describe an AI system. The availability of a low-level algorithmic description does not make a high-level description in terms of agency incorrect. To see this, first consider the case of a human being. At some level, the human organism is a biophysical system, where physical and chemical processes take place, neurons get activated, and electrochemical signals get transmitted. This may be the right level of description for some medical interventions and the neuroscientific study of brain processes. But for other purposes, it is essential to describe humans as intentional agents. It is the ascription of agency to people that allows us to explain why they vote the way they do, why they show up for work, why they do or do not keep their promises, and so on. Similarly, different levels of description are available for computational systems. At the hardware level, we may view a computer as a physical system in which electricity flows through microchips. Abstracting away from the hardware, we may also view it as executing binary logical operations, or, at an even higher level, as running software applications. Those higher levels of description are essential if we want to understand how a computer works; no software engineer could dispense with them. In the case of AI, in particular, there is not just a low level of description at which we focus on mechanisms or algorithms, but, to capture the cognitive or agentive tasks a system implements,

we may need to view the system as a goal-directed agent that responds intelligibly to its environment (see also Floridi and Sanders 2004). We may understand the system's functioning better by recognizing its high-level representations – its "beliefs" and "goals" – than by trying to unpack the precise workings of the underlying algorithms. Again, this echoes Dennett's (1987) point that taking an "intentional stance" towards a system is sometimes explanatorily necessary.

There is, in fact, a further consideration that favours the design of AI systems that qualify as intentional agents. A common criticism of AI is that the underlying mechanisms are opaque and hard to understand and predict. The quest for "explainable" AI is precisely the quest for intelligible explanations of why an AI system behaves the way it does. Explainability has been defined as "an approach focusing on providing understandable justifications for the decisions made by an AI model, aiming to answer questions like '*Why did the model make this decision?* '", where the explanation should be intelligible "from a user's perspective" (Retzlaff et al. 2024, p. 2; see also Miller 2019).[8] Explainability, so defined, may be hard to achieve at a low, algorithmic level of description, just as it is hard to explain human behaviour by looking exclusively at neurons firing while ignoring high-level cognitive and agentive functions.[9] The quest for explainability may give us reasons to design AI systems whose functioning is comprehensible in agential terms. It may prompt us to look for system architectures that support stable belief-like and goal-like states that render the system's behaviour intelligible. Questions about why a system performed certain actions are often best answered by giving high-level explanations, such as intentional ones (see also Miller 2019).[10]

4.2. Alternative possibilities

If, as I have argued, we have good reasons to view some AI systems as intentional agents, we must then ask whether we have good reasons to think that those systems have alternative possibilities to choose from. In fact, it can also be argued that once we explain an entity's behaviour by viewing it as an intentional agent, we must assume that this entity has alternative possibilities for choice (List 2019, 2023a). Although intentional explanations are not always presented in a formally precise manner, an intentional explanation of an entity's behaviour typically proceeds as follows, at least implicitly:

---

[8] Retzlaff et al. (2024) further distinguish between "post-hoc" and "ante-hoc" explainability: "[t]hey are distinguished based on whether a model is intrinsically explainable (ante-hoc), or whether explainability is achieved by [an analysis of] the model after training (post-hoc)" (p. 2).

[9] Indeed, Retzlaff et al. (2024) distinguish explainability from what they call "interpretability", which they define as "understanding the inner workings and mechanisms of an AI model, seeking to answer questions like *'How does the model work?* '", as seen "from a developer or researcher's perspective" (p. 2). The latter might focus somewhat more on low-level mechanisms.

[10] Miller (2019) notes that "it is not a stretch to assert that people will expect explanations using the same conceptual framework [namely, belief-desire-intention psychology] used to explain human behaviours. This model is particularly promising because many knowledge-based models in deliberative AI either explicitly build on such folk psychological concepts, such as belief-desire-intention (BDI) models …, or can be mapped quite easily to them; e.g. in classical-like AI planning, goals represent desires, intermediate/landmark states represent intentions, and the environment model represents beliefs" (p. 17).

1. The explanation assumes that the entity has a choice between different possible options.
2. The explanation assumes that the entity somehow considers or evaluates these options from a goal-directed perspective.[11]
3. The explanation assumes that the entity chooses one of the options on that basis.

For example, when political scientists explain why people vote for a particular party, they assume that those people face a choice between different parties, consider them based on their preferences, and make an intelligible (albeit perhaps not always rational) choice on that basis. Similarly, when economists explain consumer behaviour, they assume that consumers face choices between different consumption bundles, compare these options based on their preferences, and make a choice. The present explanatory scheme (consisting of steps 1 to 3) seems to be a common denominator of intentional explanations across a wide range of academic disciplines. And clearly, this explanatory scheme could not work without the assumption that the relevant agents face choices: alternative possibilities are a presupposition of intentional explanations (on the nature of agentive possibilities, see also Maier 2015, 2022).

This point can be reinforced by noting that intentional explanations, whether in the social sciences or in artificial intelligence, effectively have a decision-theoretic format: they attribute choice options to the agents and a mechanism of choosing one option from amongst several possible ones. Thus, if we view an AI system as an intentional agent and explain its behaviour through a decision-theoretic lens – something central to many approaches to AI, including Russell and Norvig's (2021) – we must assume that the system has alternative possibilities to choose from, just as economists or political scientists assume that human agents choose between alternative possibilities (Maier 2023; List 2023a).

Finally, the desideratum of explainability also favours systems that can be viewed as choosing between alternative possibilities. As noted by Miller (2019), good explanations in response to "why" questions are typically *contrastive*: "they are sought in response to particular counterfactual cases… That is, people do not ask why event P happened, but rather why event P happened instead of some event Q" (p. 3). For instance, we might ask: why did an agent do X rather than Y? Why did a particular self-driving car fail to stop at the junction rather than stop? Why did an autonomous trading system go ahead with a particular trade rather than refrain from doing so? Why did the diagnostic system arrive at a positive rather than negative diagnosis on some case? And so on. By attributing alternative possibilities to a system among which this system makes a choice, intentional explanations have the desired contrastive format.

4.3. Causal control

The third question we must ask is whether we have good reasons to think that AI systems have causal control over their actions. It is first helpful to clarify how one would understand "causal control" in the case of a human being. The key question there is whether the person's high-level mental states, such as the intention to perform the action, make the right causal difference to that action such that the action is not exclusively explained by physical states of the brain

[11] "Consideration" or "evaluation" could be anything ranging from slow and deliberative to fast and instinctive.

and body, like a bodily reflex (Woodward 2008, List and Menzies 2009, Raatikainen 2010). For a mental state to be a *difference-making cause* of an action, in turn, two counterfactuals must be true:

1. If the person did not have that mental state, they would not perform the action.
2. If the person had that mental state in other similar circumstances, they would also perform the action.

For example, my intention to vote "yes" in a committee (a mental state) is the difference-making cause of the act of raising my arm at the right moment. If I didn't have that intention, I wouldn't raise my arm; and if I had the intention in other similar circumstances, I would also raise my arm. Thus a mental state is a difference-making cause of an action if the performance of the action systematically co-varies with the presence or absence of that mental state, holding other things fixed (List and Menzies 2009). The difference-making account of causal control can be spelt out further, but what matters for present purposes is that, in the case of genuine actions as opposed to mere bodily processes like digestion or reflexes, we cite mental states as the explanatorily significant difference-makers. Note that citing a mental state as a difference-making cause of an action is consistent with recognizing that this mental state is implemented by physical states of the brain and body.

A similar analysis can be given by introducing the notion of a "control variable" for some outcome. Control variables are "parameters which, when changed, lead to systematic changes in other variables of interest" (Roskies 2012, p. 329). When we explain human actions, as Campbell (2010, p. 26) notes,

"(a) psychological variables [such as mental states] function as control variables for the outcomes in which we are interested,
(b) what is going on at a psychological level of description supervenes on [is implemented by] what is going on at a physical level of description, but
(c) at the physical level, there are no control variables for the outcomes in which we are interested."

Physical-level states, such as highly specific neural states of the brain, are too fine-grained to serve as control variables for human actions. Recall again that we would explain the raising of my arm when a vote is taken by citing my voting intention rather than a particular microstate of my brain and body. Indeed, influencing human behaviour at the agential level, by providing people with information and motivations, is usually more effective than influencing it at a purely physical level, by trying to influence brain activity directly.

Now it should be clear what it would take for an AI system to have causal control over its actions. The key question is whether, to explain the system's actions, it is always better – more empirically adequate, more informative, more parsimonious – to cite low-level microstates of the system as causes, for instance descriptions at the level of the underlying algorithm, or whether it is sometimes better to cite high-level representational or goal states. Equivalently, we must ask whether the control variables for the system's actions are always to be found at a

low algorithmic level or at least sometimes at a higher representational level. In the latter case, our explanations of what the AI system does would refer to the analogues of mental rather than physical causes. The system could then be said to have causal control over its actions.

Again, the quest for explainable AI is relevant. An AI system whose actions systematically co-vary with its representational and goal states will be more explainable than one whose behaviour can be viewed only as the opaque result of low-level algorithmic processes. Explainability thus gives us a reason to design AI systems that meet the condition of causal control.

In sum, to the extent that some AI systems are best explained as intentional agents with alternative possibilities to choose from and causal control over their actions, those systems may be said to have free will, at least under the present, pragmatic definition. Interestingly, Floridi and Sanders (2004, p. 349) share the view that AI systems can qualify as agents in a high-level sense (noting that "[a]genthood … depends on a LoA [level of analysis]"), but they suggest that the relevant concept of agency might be one "not necessarily exhibiting free will, mental states or responsibility". Yet they seem to assume that free will and "mindedness" require "some special internal states, enjoyed only by human and perhaps super-human beings" (p. 366). As noted earlier, AI systems are unlikely to have such special internal states. However, once we reconceptualize agency and free will in the present, less metaphysically demanding way, there may be more common ground between their view and mine. Floridi and Sanders concede that the relevant artificial agents "are already free in the sense of being non-deterministic systems" (p. 366) (which I would understand as "non-deterministic at the relevant level of analysis", as discussed in section 5.2 below), and they further say: "the agents in question satisfy the usual practical counterfactual: they could have acted differently had they chosen differently, and they could have chosen differently because they are interactive, informed, autonomous and adaptive" (p. 366). So, once the checklist for free will is defined as suggested here, Floridi and Sanders's position is less far from mine that it might initially seem.

5. Further questions

My argument for the possibility of free will in AI invites several further questions.[12]

5.1. If a computer implements an AI system that satisfies the above-mentioned conditions for free will, which entity is the bearer of free will: the computer, the software, or something else?

It seems counterintuitive to suggest that a computer or a smartphone could acquire free will simply by running an appropriate AI app. On my account, however, free will is a high-level property of the AI system in its entirety, not a property of the underlying physical device. So, free will would be a system-level property of the choice-making agent that is being *implemented* by the computer with the relevant software. Free will is not a property of the computer *qua* physical device. In the same way, one would say that a human being has free will *qua* intentional agent, and not that the underlying brain has free will *qua* physical organ. The brain is part of the hardware that *implements* the high-level system with free will.

[12] This section has particularly benefitted from Sven Nyholm's helpful comments and suggestions.

5.2. Doesn't the fact that AI systems are based on deterministic algorithms rule out free will from the outset?

The first thing to note is that, according to the widely held "compatibilist" view in philosophy, free will is compatible with determinism, whether it is determinism in the brain and body or determinism in the physical world as a whole. Dennett (1984, 2003) is one among many philosophers who hold such a view. In a recent survey of professional philosophers, almost 60% of respondents described themselves as compatibilists (Bourget and Chalmers 2023). If compatibilism is correct, the mere fact that an entity's "hardware" is deterministic would not exclude the possibility that the entity has free will. This point applies not only to human beings but arguably also to AI systems (Maier 2023).

However, since I have included alternative possibilities in the definition of free will, I cannot appeal to those versions of compatibilism that drop the alternative-possibilities requirement for free will. Recall that Dennett, for instance, drops this requirement in some of his writings (e.g., Dennett 1984). If we retain that requirement, then it is not clear how a system that is based on deterministic algorithms could make real choices between alternative possibilities. At any time, the system's state would fully determine what the system does next; the system would never face a genuine fork in the road, where it could do one thing or another.

However, as already noted, some compatibilists – including Dennett in other writings (e.g., Dennett 2003 and Taylor and Dennett 2002) – have offered strategies for (re)defining the notion of "having alternative courses of action" such that it becomes compatible with determinism. To explain my own preferred strategy, applied to the case of AI, let me begin by recalling that there are different levels at which we can describe an AI system: a micro-level at which we refer to (gazillions of) binary operations in logic gates, and a macro-level at which we refer to the cognitive and agentive processes realized. Micro-level descriptions are more fine-grained, macro-level ones more coarse-grained. Once we recognize that there are different such levels of description, we can also see that there are different notions of "possibility", which are relevant at different levels. At the micro-level the relevant notion is *possibility, conditional on the system's micro-state*, while at the macro-level it is *possibility, conditional on the system's macro-state*. The latter (macro-level) notion is less constrained than the former (micro-level) one, insofar as it conditionalizes on a coarse-grained macro-level state instead of a fine-grained micro-level state. This, in turn, implies that the more permissive, macro-level notion of possibility may admit alternative possibilities even when the more restrictive, micro-level notion doesn't. Consequently, our best *macro-level* explanations of a system may describe that system as indeterministic, even if its underlying *micro-level* processes are deterministic. In fact, this point holds generally for systems that can be described at different levels (Butterfield 2012, Yoshimi 2012, and List 2014): the determinism/indeterminism distinction is not preserved under changes in the level of description that we use to analyze a system.

On this account, what matters for alternative possibilities in any system – human, animal, AI – is whether the system is *best explained* by depicting it as an intentional agent capable of choosing between alternative possibilities. If the answer is "yes", then the fact that, at a lower

level, there are deterministic processes is irrelevant. I propose that "alternative possibilities" in the context of free will should be understood as alternative possibilities at the (macro-)level of agency (List 2014, 2023a). And as already argued, we have good explanatory reasons for attributing such alternative possibilities to AI systems, insofar as they qualify as choice-making agents. (The claim that there is a distinct "agentive" notion of possibility has also been defended by Maier 2015, 2022.)

It is important to note that although free will requires alternative possibilities at the level of agency, free will is nonetheless consistent with predictability: an entity can qualify as having free will – as being an agent capable of choosing and controlling its actions – while being predictable in its behaviour, for instance because the entity makes its choices for intelligible reasons. Human beings are often quite predictable, but this is not because they lack free will but because they make their choices based on reasons that a predictor may understand. For example, I disprefer alcoholic drinks and therefore choose a non-alcoholic drink each time I go to a restaurant. My friends, who know my preferences, can predict those choices, but that doesn't mean that I do not make genuine choices in the first place or that I lack the capacity to choose otherwise. Even if I predictably choose the non-alcoholic drink, the choice is mine.

5.3. Wouldn't the present analysis have the counterintuitive implication that even simple optimizing algorithms have free will?

Consider a chess-playing computer. For each possible configuration of the chessboard, the system considers all possible moves permitted by the rules of chess and chooses the move deemed best by some objective function, which encodes the algorithm's "goals". This description has a decision-theoretic format and might thus suggest that our chess computer is a (simple) choice-making agent and thereby in principle the sort of entity that has free will. This conclusion is counterintuitive, since we could equally explain the chess computer in non-agential terms. We can think of it as a deterministic system whose possible states are the possible configurations of the chessboard and whose state change rule is a deterministic function mapping each state to a unique next state, namely precisely the one that, under the earlier, choice-theoretic description, would have been described as "maximizing the value of the relevant objective function". This redescription makes no reference to agency, choice, or alternative possibilities and seems to explain the system's behaviour equally adequately.[13] A similar point could be made about other entities that admit both non-intentional and intentional explanations, such as a thermostat. A thermostat can be viewed as a mechanical device, but also as a rudimentary agent that "chooses" between activating and de-activating the heating, depending on whether it "believes" the actual temperature is too low or too high relative to some "desired" target temperature (see also Dennett 1987).

I suggest that we should *not* attribute free will to a system unless viewing that system as a choice-making agent is explanatorily superior to not doing so. Since non-agential explanations

[13] However, one might argue that *if* we need to refer to the original objective function to define the state change rule, we haven't genuinely eliminated the choice-theoretic format. On formal differences between intentional and non-intentional explanations, see also Orseau, McGregor McGill, and Legg (2018).

are perfectly feasible for the chess computer and the thermostat, and by some standards simpler, it is unnecessary to view those systems as choice-making agents. It would be an over-interpretation to ascribe free will to them; we would be ascribing to them richer cognitive capacities than they plausibly have. By contrast, in the human case, the ascription of choice-making agency is often explanatorily indispensable, and so the attribution of free will seems warranted. If an AI system is so complex as to render non-intentional explanations practically infeasible, then the same could be said about such a system.

5.4. Isn't the claim that AI systems can have free will challenged by the fact that many such systems do not take any initiatives by themselves and act only when prompted to do so?

Present AI systems are often only very reactive agents: they take actions only in response to human prompts, and once they have completed any given task, they remain passive and resume their activity only when prompted again. A self-driving car, for instance, does nothing until instructed to drive to a particular destination. Similarly, many LLM systems produce outputs only in response to specific prompts. In light of this, Nyholm (2018, p. 1201) has argued that "we ought not to regard [AI systems] as acting on their own, independently of any human beings. Rather, the right way to understand the agency exercised by these machines is in terms of human–robot collaborations, where the humans involved initiate, supervise, and manage the agency of their robotic collaborators."

My claim, however, is only that whenever a system is in a phase of choice-making agency, it exhibits a form of free will, by satisfying the three above-mentioned conditions. Perhaps some systems go into such a phase only when activated by certain prompts and remain "on standby" for the rest of the time, so that their agency becomes periodically inactive, a bit like a hibernating animal whose agency is temporarily dormant.

Furthermore, we can easily imagine AI systems that take on temporally extended tasks involving long phases of choice-making agency. Imagine an autonomous military drone that is tasked with monotoring a coastline on a long-term basis and that is capable of evaluating and flexibly responding to various threat situations, including ones that aren't predefined. One can think of such a system as capable of taking initiatives. Similarly, robotic pets may be designed to be spontaneous and to take initiatives, while pursuing longer-term goals, such as companionship with a human being. The extent to which a system has the capacity to take initiatives lies on a continuum and depends on how complex, flexible, and long-term its objectives are.

5.5. If an AI system has free will, does this imply that the system is also conscious?

While some researchers (such as Tononi et al. 2022) see free will and consciousness as being connected, the mainstream approach in philosophy is to treat them as separate. Conceptually, as I see it, consciousness is neither necessary nor sufficient for free will. Free will, as defined here, requires intentional agency, alternative possibilities, and causal control over one's actions. Consciousness requires the presence of subjective experiences that one undergoes from

a first-person perspective: there is something it is like to be a conscious subject, for that subject, as Thomas Nagel (1974) famously put it. Free will can be understood as a third-personal notion, while consciousness is inherently first-personal. On the present picture, a system could in principle qualify as having intentional agency, alternative possibilities, and causal control over its actions – and thereby as having free will – without subjectively experiencing anything. An example of such a system would be a "philosophical zombie", which is considered by many to be a logically coherent, albeit purely hypothetical scenario (Chalmers 1996; Dennett disagrees; see, e.g., Dennett 2005). Conversely, a system could in principle have subjective experiences without actively choosing and controlling anything. A hypothetical example could be some kind of passive "experience machine".

Empirically, of course, paradigmatic agents with free will, such as human beings or other complex animals, may also be conscious, and vice versa; Dennett would presumably agree with this claim. If that is right, however, it would establish a contingent, rather than conceptually necessary connection between free will and consciousness, especially in the biological world. A conceptually necessary connection would be established only if one could show that intentional agency itself requires consciousness. According to some philosophical views, consciousness is indeed necessary for agency and/or intentionality, but my preferred methodology is to keep our theories as modular as possible and to use "thin" definitions of key concepts, which do not rely on too many built-in assumptions. For this reason, I here understand free will as a third-personally describable phenomenon, and I do not assume that free agents must also be conscious (even though, in practice, many or most of them are).

Needless to say, free will should not be confused with the *experience* of free will. One could have that experience without really having free will (for instance, if free will were an illusion). And if there could be free will in a non-conscious entity, then there could also be free will without any experience of it. Recall once more that I have defined free will in a pragmatic and metaphysically undemanding way, and so it is not out of the question that some AI systems could qualify as having free will without having conscious experiences.

5.6. Does free will in AI imply that AI systems are capable of bearing moral responsibility?

We must distinguish between free will and moral responsibility. The former, as defined here, is primarily a descriptive and explanatory notion; the latter is normative. I have defined free will as the capacity for intentional agency, for choice between alternative possibilities, and for causal control over the resulting actions. This three-part capacity is what makes the explanatory logic of choice-making agency applicable to an entity. Free will, so understood, need not be only human, but could be present in non-human animals too. By contrast, the capacity to bear moral responsibility requires a richer form of agency, namely moral agency, which includes the capacity for moral cognition. Non-human animals arguably lack that capacity, despite having the sort of agency required for bare free will. Free will is thus necessary but not sufficient for the capacity to bear moral responsibility. That said, the quest for *ethical* AI might be viewed as the quest for designing artificial *moral* agents, and AI systems with free will may eventually become candidates for the ascription of moral responsibility.

6. Concluding remarks

To determine whether an AI system has free will, we should not be asking: does the system exhibit some mysterious property, is it unpredictable, or are its algorithms indeterministic? Rather, we should be asking: is the system *best explained* as an intentional agent, with the capacity for choice between alternative possibilities, and causal control over its actions? If we have good grounds for applying the explanatory logic of choice-making agency to a system, it may be said to have free will in a practically relevant, non-mysterious sense. The system will then arguably have the kind of "free will worth wanting", as Dennett (1984) calls it, and it will satisfy a key necessary condition for bearing moral responsibility. Anyone who considers it desirable for AI systems to function as moral agents should find this conclusion congenial.

My argument for the possibility of free will in AI resembles a similar argument in the case of another class of artificial agents: corporations and other organized collectives. The claim that such entities constitute intentional agents can also be defended on interpretivist grounds, inspired by Dennett (see Tollefsen 2015), or on functionalist grounds, by noting that suitably organized collectives satisfy the functional conditions for belief-desire agency, albeit based on a social as opposed to electronic hardware (List and Pettit 2011, List 2021). Indeed, social scientists commonly explain the behaviour of such entities through the lens of decision or game theory, by attributing to them the ability to choose between different options in a goal-directed and strategically rational manner. For example, the theory of the firm in economics represents firms and corporations as rational profit-maximizing agents.

The recognition that some collectives constitute choice-making agents in their own right raises the issue of free will too. At first, one might think that "corporations (and other highly organized collectives like colleges, governments, and the military) are effectively puppets, dancing on strings controlled by external forces", as Kendy Hess (2014, p. 241) notes. However, Hess argues that once we properly recognize how corporate agents function, we have good reasons to think that they "act from their own 'actional springs' … and from their own reasons-responsive mechanisms" (ibid.).[14] This, for Hess, supports the claim that "they act freely and are morally responsible for what they do" (ibid.).

Furthermore, one may reach this conclusion not only by reference to the criterion of reasons-responsiveness, as invoked by Hess, but also by reference to the checklist of conditions used here: intentional agency, the capacity for choice between alternative possibilities, and control over the resulting actions. Corporations and other suitably organized groups agents arguably satisfy these conditions too (List 2023b). The argument for free will in corporate entities is therefore similar to the one in the case of AI.

Free will, we may conclude, is not restricted to human beings and other complex animals, but can in principle occur in non-biological agents too. Group agents and AI systems are two different examples of a similar phenomenon.

[14] Hess attributes the notion of "acting from one's own actional springs" to Haji (2006).

References

Ananthaswamy, A. (2024). "How close is AI to human-level intelligence." *Nature* 636: 22–25.

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.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Blum, M., and L. Blum (2022). "A Theoretical Computer Science Perspective on Free Will." https://arxiv.org/abs/2206.13942

Bourget, D., and D. Chalmers (2023). "Philosophers on Philosophy: The 2020 Philpapers Survey." *Philosophers' Imprint* 23(11).

Bratman, M. E. (1987). *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Butterfield, J. (2012). "Laws, Causation and Dynamics at Different Levels." *Interface Focus* 2(1): 101–114.

Campbell, J. (2010). "Control Variables and Mental Causation." *Proceedings of the Aristotelian Society* 110: 15–30.

Carstensen, M. W., S. Sellmaier, P. C. J. Taylor, and O. Deroy (2023). "Dualists and physicalists agree, free will is incompatible with determinism." *Philosophical Psychology*. https://doi.org/10.1080/09515089.2023.2263044

Chalmers, D. (1996). *The Conscious Mind*. New York: Oxford University Press.

Chalmers, D. (2023). "Could a Large Language Model Be Conscious?" *Boston Review*. https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/

Delcker, J. (2018). "Europe divided over robot 'personhood'." *Politico*. https://www.politico.eu/article/europe-divided-over-robot-ai-artificial-intelligence-personhood/

Dennett, D. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting.* Cambridge, MA: MIT Press.

Dennett, D. (1987). *The Intentional Stance.* Cambridge, MA: MIT Press.

Dennett, D. (2003). *Freedom Evolves.* London: Penguin.

Dennett, D. (2005). *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*. Cambridge, MA: MIT Press.

Dennett, D. (2009). "Intentional Systems Theory." In *The Oxford Handbook of Philosophy of Mind*, edited by A. Beckermann, B. P. McLaughlin, and S. Walter, 339–350. Oxford: Oxford University Press.

Dubber, M. D., F. Pasquale, and S. Das (eds.) (2020). *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press.

Farnsworth, K. D. (2017). "Can a Robot Have Free Will?" *Entropy* 19(5), 237.

Floridi, L. (2023). *The Ethics of Artificial Intelligence*. Oxford: Oxford University Press.

Floridi, L., and J. W. Sanders (2004). "On the Morality of Artificial Agents." *Minds and Machines* 14(3): 349–379.

Frankfurt, H. (1969). "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66(23): 829–839.

Fossa, F. (2018). "Artificial moral agents: Moral mentors or sensible tools?" *Ethics and Information Technology* 20(2): 115–126.

Gurnee, W., and M. Tegmark (2023). "Language Models Represent Space and Time." https://arxiv.org/abs/2310.02207

Haji, I. (2006). "On the Ultimate Responsibility of Collectives." *Midwest Studies in Philosophy* 30(1): 292–308.

Harris, S. (2012). *Free Will.* New York: Simon and Schuster.

Hess, K. (2014). "The Free Will of Corporations (and Other Collectives)." *Philosophical Studies* 168(1): 241–260.

Kane, R. (ed.) (2002). *The Oxford Handbook of Free Will*. Oxford: Oxford University Press.

Krausová, A. and H. Hazan (2013). "Creating Free Will in Artificial Intelligence." In *Proceedings of the International Conference Beyond AI 2013*, Pilsen, Czech Republic, edited by J. Romportl et. al. https://www.beyondai.zcu.cz/files/BAI2013_proceedings.pdf

Levinstein, B. A., and D. A. Herrmann (2024). "Still no lie detector for language models: probing empirical and conceptual roadblocks." *Philosophical Studies*. Online early, https://doi.org/10.1007/s11098-023-02094-3

List, C. (2014). "Free will, determinism, and the possibility of doing otherwise." *Noûs* 48(1): 156–178.

List, C. (2019). *Why Free Will is Real*. Cambridge, MA: Harvard University Press.

List, C. (2021). "Group Agency and Artificial Intelligence." *Philosophy and Technology* 34(4): 1213–1242.

List, C. (2023a). "Agential Possibilities." *Possibility Studies and Society* 1(4): 1–10.

List, C. (2023b). "Do group agents have free will?" *Inquiry*. https://doi.org/10.1080/0020174X.2023.2218721

List, C., and P. Menzies (2009). "Non-reductive Physicalism and the Limits of the Exclusion Principle." *Journal of Philosophy* 106(9): 475–502.

List, C., and P. Pettit (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.

Maier, J. (2015). "The Agentive Modalities." *Philosophy and Phenomenological Research* 90(1): 113–134.

Maier, J. (2022). *Options and Agency*. Heidelberg: Springer.

Maier, J. (2023). "Artificial Intelligence and Free Will". https://pub.towardsai.net/artificial-intelligence-and-free-will-27e157437e58.

Matthias, A. (2004). "The responsibility gap: Ascribing responsibility for the actions of learning automata." *Ethics and Information Technology* 6(3): 175–183.

Miller, T. (2019). "Explanation in artificial intelligence: Insights from the social sciences." *Artificial Intelligence* 267: 1–38.

Moore, M. S. (2020). *Mechanical Choices: The Responsibility of the Human Machine*. Oxford: Oxford University Press.

Nagel, T. (1974). "What Is It Like to Be a Bat?" *The Philosophical Review* 83: 435–450.

Nyholm, S. (2018). “Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci.” *Science and Engineering Ethics* 24: 1201–1219.

Nyholm, S. (2020). *Humans and Robots: Ethics, Agency, and Anthropomorphism*. London: Rowman & Littlefield.

O’Connor, T. (2010). “Free Will.” *The Stanford Encyclopedia of Philosophy* (Winter 2010 Edition), edited by E. N. Zalta and U. Nodelman. https://plato.stanford.edu/archives/win2010/entries/freewill/

Orseau, L., S. McGregor McGill, and S. Legg (2018). “Agents and Devices: A Relative Definition of Agency.” https://arxiv.org/abs/1805.12387

Pereboom, D. (2001). *Living without Free Will.* Cambridge: Cambridge University Press.

Raatikainen, P. (2010). “Causation, Exclusion, and the Special Sciences.” *Erkenntnis* 73(3): 349–363.

Retzlaff, C. O., A. Angerschmid, A. Saranti, D. Schneeberger, R. Röttger, H. Müller, and A. Holzinger (2024). “Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists.” *Cognitive Systems Research* 86(101243): 1–17.

Roskies, A. L. (2012). “Don’t Panic: Self-Authorship without Obscure Metaphysics.” *Philosophical Perspectives* 26: 323–342.

Russell, S., and P. Norvig (2021). *Artificial Intelligence: A Modern Approach*. 4th Edition. Harlow: Pearson Education.

Sanchis, E. (2018). “A Model of Free Will for Artificial Entities.” https://arxiv.org/abs/1802.09317

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

Solum, L. B. (1992). “Legal personhood for artificial intelligences.” *North Carolina Law Review* 70(4): 1231–1287.

Taylor, C., and D. Dennett (2002). “Who’s Afraid of Determinism? Rethinking Causes and Possibilities.” In *The Oxford Handbook of Free Will*, edited by R. Kane, 257–277. Oxford: Oxford University Press.

Tononi, G., L. Albantakis, M. Boly, C. Cirelli, and C. Koch (2022). “Only what exists can cause: An intrinsic view of free will.” https://arxiv.org/abs/2206.02069

Tollefsen, D. P. (2015). *Groups as Agents*. Cambridge: Polity.

Woodward, J. (2008). “Mental Causation and Neural Mechanisms.” In *Being Reduced: New Essays on Reduction, Explanation, and Causation*, edited by J. Hohwy, and J. Kallestrup, 218–262. Oxford: Oxford University Press.

y Arcas, B. A., and P. Norvig (2024). “Artificial General Intelligence Is Already Here.” https://www.noemamag.com/artificial-general-intelligence-is-already-here/

Yoshimi, J. (2012). “Supervenience, Dynamical Systems Theory, and Non-Reductive Physicalism.” *The British Journal for the Philosophy of Science* 63(2): 373–398.