Can AI systems have free will?

Christian List[*]

Discussion paper, November/December 2024

Abstract: While there has been much discussion of whether AI systems could function as moral agents or acquire sentience, there has been relatively little discussion of whether AI systems could have free will. In this article, I sketch a framework for thinking about this question. 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 it is explanatorily indispensable 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

We are witnessing an artificial intelligence revolution. AI systems are becoming ubiquitous in many areas: from transportation to business, finance, public administration, science, medicine, the law, communication, entertainment, and the military. This trend raises many questions. Are AI systems safe? Will they behave ethically? Who is responsible when something goes wrong? Could AI systems become conscious? Should society even assign legal personhood to them?[1]

Despite the explosion of work on these questions, one question has received surprisingly little attention: can AI systems have free will? Free will is associated with autonomous agency and is a precondition for moral responsibility. Since AI systems increasingly take on decision-making tasks with high stakes, people ask: could it make sense to attribute responsibility to AI systems themselves, not just to their manufacturers, owners, and operators? Could we blame those systems as agents in their own right? The issue of free will matters for all these questions.

In this article, I will describe a framework for thinking about free will in AI. My aim is to lay out what it would take for an AI system to have free will. I will consider a checklist of conditions for free will (building on List 2019) and suggest that free will in AI is less far-fetched than perhaps expected.

Among the few prior works discussing free will in AI, most employ more demanding conditions for free will and thus typically reach more negative conclusions (see, e.g., Floridi and Sanders 2004, Krausová and Hazan 2013, Farnsworth 2017, and Sanchis 2018).[2] Others

[*] I am very grateful to Sven Nyholm and 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 explained in section 5.2 below, I do not agree that unpredictability is required (or even congenial) for free will.

focus on the public's *beliefs* about whether AI systems have free will (Astobiza 2024), also suggesting a broadly negative answer. The closest precursor to the present approach can be found in a blog post by Maier (2023), who, like me, argues for the possibility of free will in AI by emphasizing that AI systems can be viewed as choice-making agents to which decision-theoretic models are applicable. I will develop this idea by reference to the account of free will presented in List (2019), where the application to AI was already briefly anticipated. The present checklist for free will should be diagnostically useful, among other things by capturing key features of the commonsense notion of free will that is often considered necessary (albeit not by itself sufficient) for moral responsibility.

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 their influential textbook, Stuart Russell and Peter Norvig (2021, p. 7) characterize the field of AI as "the study of agents that receive percepts from the environment and perform actions".

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 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. Weak AI is already common. Think of chess-playing computers or smart route planners. Strong AI is increasingly becoming a reality, as exemplified by generative AI chatbots such as ChatGPT, which 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] In any case, current AI systems already 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 military contexts.

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. "Generative AI", the currently dominant approach, refers to the implementation of 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. Google characterizes AI as "a set

[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.

of technologies that are based primarily on machine learning and deep learning, used for data analytics, predictions and forecasting, object categorization, natural language processing, recommendations, intelligent data retrieval, and more".[4] 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, on a first characterization, is an entity's capacity to choose and control its own actions (for an overview, see Kane 2011). According to our commonsense understanding, we human beings have free will. 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 choices that bypass the laws of nature, 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 implausible. However, 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 kind 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 should support those distinctions and not rule out free will from the outset.

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, 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, such as 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.

[4] See https://cloud.google.com/learn/what-is-artificial-intelligence, accessed on 19 October 2024.

I will assume that these three conditions are jointly necessary and sufficient for free will. The distinctions drawn by judges illustrate them. Someone in full possession of their cognitive and agentive capacities appears to meet 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 think that alternative possibilities are not needed for free will (see the review in Kane 2011). Many so-called compatibilists, in particular, think 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" or "owner". Others, especially so-called libertarians, think that dropping alternative possibilities would be a watering-down of the idea of free will, and some other compatibilists agree and seek to come up with a determinism-friendly way of redefining the notion of "having alternative courses of action" (for a discussion of the available strategies, see, e.g., List 2014). Conventionally, we think that free choices require a "fork in the road" ahead of us, where we could do one thing or another. To capture this, I will here include alternative possibilities in the definition of free will.

Free-will sceptics claim that people never have the three capacities required for free will. What we conventionally consider a free choice, they 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).

This article is not the place to respond to such free-will scepticism in general. Note, however, that society, including our legal system, operates on the assumption that human beings normally have the three capacities required for free will: agency, choice, and control over their actions. The sciences of human behaviour operate on this assumption, too (List 2019, 2024). Disciplines ranging from economics and political science to anthropology and history all assume that *the people whose behaviour they study are intentional agents who make choices and have a relevant form of causal control over those choices*. Those sciences thereby assume – albeit often implicitly – that people have free will under the present, pragmatic 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 no longer be talking about agents making choices. Rather, we would have to reconceptualize people as passive spectators: organisms that are moved by factors beyond their control, in roughly the same way in which 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 some support to free will as a working hypothesis.

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; but so would humans. On my analysis, we should not look for some mysterious property, but use the checklist of the three above-stated conditions to assess the system. An AI system would thus have free will *if and only if* it has intentional agency, alternative possibilities to choose from, and causal control over its actions (List 2019).

To determine whether a system fulfills those conditions, moreover, I propose that we should not look for anything mysterious either, but ask whether, to understand and predict the system's behaviour, it is *explanatorily indispensable* to view the system as (i) an intentional agent with (ii) alternative possibilities 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 may 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 (or utility assignments to) different such states. Given these definitions, AI systems can plausibly qualify as agents with beliefs and desires. Recall Russell and Norvig's (2021) characterization of AI in terms of agency. Indeed, the agentive and cognitive capacities of AI systems have recently advanced dramatically, and more and more such systems arguably qualify as (sometimes even relatively complex) belief-desire agents (Bratman 1987).

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 such as large language models cannot generally be interpreted as accurate reflections of stable internal states that play a belief role. For example, a system may give inconsistent responses to different users or in response to different prompts, and it may be best viewed as a "stochastic parrot" (Bender et al. 2021) producing an output that is likely to be a good fit for the given input, relative to some loss function. This would speak against the view that the system genuinely has the kinds of belief-and-desire states that would be 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 be that, as a byproduct of this minimization exercise, some internal states of the system come to play a representational (belief) role. 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, it is easy to identify problem-solving tasks in which an AI system would benefit from having internal states that play a belief role. So, I agree with Levinstein and Herrmann that for an AI system, like for us, "there are plenty of contexts in which it is very useful to have an accurate map of the world, in order to guide action" (p. 8), and so it is "largely an empirical matter" (ibid.) whether AI systems have beliefs and, I would add, other stable intentional states.

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 agency. However, this criticism conflates different levels at which we might describe an AI system. 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 a human being can also be described as an intentional agent. As noted, the agential level of description is essential for the sciences of human behaviour, as well as for everyday interactions. It is often explanatorily necessary to view people as intentional agents and to give intentional and not just dynamic or stochastic explanations of their behaviour. Dennett (1987) calls this "the intentional stance". It is precisely the ascription of intentional states to people that allows us to explain why people vote the way they do, why they show up for work every day, 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. More abstractly, we may view it as executing binary logical operations. And at an even higher level, we may view it as running software applications. Far from being just a heuristic, the software level of description is essential if we want to understand how the computer works; no software engineer could dispense with it.

Arguably, something similar is true for many AI systems. There is a low level of description at which we focus on mechanisms or algorithms. But we may also describe the cognitive or agentive tasks that a system implements, and this may require us to view the system as a goal-directed agent that responds intelligibly to its environment (see also Floridi and Sanders 2004). We may better understand the system's behavioural regularities by recognizing its goal-directed agency than by trying to unpack the detailed workings of the underlying algorithm. We cannot generally dispense with the agential level of description.

A common criticism of AI is that the underlying mechanisms are opaque and hard to understand and predict. The quest for "explainable" AI is the quest for intelligible explanations of why an AI system behaves the way it does. Explainability 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. The quest for explainability may give us reasons to *design* AI systems whose high-level functioning is comprehensible in agential terms. It may prompt us to look for stable belief-like and goal-like states of the system that render its behaviour intelligible.

4.2. Alternative possibilities

If we view some AI systems as intentional agents, we must then ask whether those systems have alternative possibilities to choose from. It can 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). Note that intentional explanations of an entity's behaviour fit the following three-part structure:

1. The explanation assumes that the entity has a choice between different possible options.
2. The explanation assumes that the entity considers or evaluates these options, whether in a fast and "instinctive" manner or in a slow and "deliberative" one.
3. The explanation assumes that the entity chooses one option 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. 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, 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 2024).

4.3. Causal control?

Finally, could an AI system qualify as having causal control over its actions? It is 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 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 didn't have that mental state, they would not perform the action.

2. If the person had that mental state in other similar circumstances, they would still 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 similar circumstances, I would still 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. 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.

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. Others have made this point 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 intentional 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 precise 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 typically 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 informative, more parsimonious – to cite low-level microstates of the system as causes, for instance state descriptions at the level of the underlying algorithm, or whether it is sometimes better to cite high-level representational or goal states. In the latter case, our explanations of what the AI system does would refer to the analogues of high-level mental rather than low-level 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 high-level representational and goal states is likely to 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 "[a]genthood … depends on a LoA [level of analysis]" and that an AI system can qualify as an agent in a high-level sense, but they suggest that the relevant concept of agency might be one "not necessarily exhibiting free will, mental states or responsibility". However, they seem to reach this conclusion because they think 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. Once we reconceptualize free will in the present, less metaphysically demanding way, there may be more common ground between their view and the one defended here. 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"; see section 5.2 below). 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.

The 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 well-organized group agents. Those entities are also artificial agents that display autonomous choice-making agency, albeit based on a social as opposed to electronic hardware. Social scientists commonly explain the behaviour of such entities through the lens of decision theory or game theory, by attributing to them the ability to choose between different options in a goal-directed and often strategically rational manner. An example can be found in the theory of the firm in economics, which represents firms and corporations as rational profit-maximizing agents. At first, one may 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, but she argues that once we consider how corporate agents function, we have reason to think that they "act from their own 'actional springs' … and from their own reasons-responsive mechanisms".[5] Their behaviour can be highly rational and responsive to reasons internal to the organization. This, for Hess, supports the claim that "they act freely and are morally responsible for what they do" (ibid.). Furthermore, as argued in List (2023b), one may reach this conclusion not only by reference to the criterion of reasons-responsiveness, which is used by Hess, but also by reference to the checklist employed in this paper. That is, corporations and other suitably organized groups agents display intentional agency, the

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

capacity for choice between alternative possibilities, and control over the resulting actions. The argument for free will in corporate entities is therefore very 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 may be examples.

5. Further questions

My argument for the possibility of free will in AI invites several further questions.[6]

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 may seem counterintuitive to say that a computer or even a smartphone acquires free will simply by running an 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, 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. This analysis seems natural if we apply it to an "embodied" AI system, such as a robot or a self-driving car, which satisfies the relevant conditions. We would then say that the system as a whole has free will.

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, either in the brain and body or in the physical world as a whole. In the most recent Philpapers survey of professional philosophers, for example, 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 automatically exclude the possibility that the entity has free will. This point in principle carries over to AI too (see also Maier 2023).

However, since I have explicitly 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 compatibilists are divided between those who give up that requirement and those who try to redefine it in a determinism-friendly manner.) At first, this appears to limit the room for defending free will in AI significantly. In particular, it may seem that if a system is based on deterministic algorithms, it could never make real choices between alternative possibilities. At any point in time, the system's state would fully determine

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

what the system does next; the system would never face a genuine fork in the road, where it could do one thing or another.

But, as emphasized, we must distinguish between different levels at which we can describe the 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. Furthermore, any talk of "possibility" is level-specific too. 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 notion may admit alternative possibilities even when the former doesn't, and so our best macro-level explanations of a system may describe it as indeterministic, even if its micro-level processes are deterministic (this point holds generally for systems that can be described at different levels; see Butterfield 2012, Yoshimi 2012, and List 2014). What matters for alternative possibilities in an AI system 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 algorithmic processes is irrelevant. I propose that "alternative possibilities" in the context of free will is best understood as alternative possibilities at the (macro-)level of agency (List 2014).

Furthermore, free will is 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 agents are often quite predictable, but not because they lack free will but because they make their choices based on reasons that a predictor may understand. I may disprefer alcoholic drinks and therefore choose a non-alcoholic drink each time I go to a bar or restaurant. My friends, who know my preferences, are able to predict those choices, but that doesn't mean that I don't 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 interpreted as "maximizing the value of the relevant objective function". This dynamic-system redescription makes no reference to agency, choice, or alternative possibilities and seems to explain the system's behaviour equally adequately.[7] A similar point could be made about other entities that admit both dynamical 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 (see also Dennett 1987). I suggest, however, that we should not attribute free will to a system *unless viewing the relevant system as a choice-making agent is explanatorily indispensable*. Since non-agential explanations are perfectly feasible for the chess computer and the thermostat, it is unnecessary to view those systems as choice-making agents. It would be an over-interpretation to ascribe free will to them. By contrast, in the human case, the ascription of choice-making agency is often explanatorily indispensable, and so the ascription of free will seems warranted. If an AI system is so complex as to rule out purely dynamical and non-intentional explanations, 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 "dormant" for the rest of the time, so that their agency becomes periodically inactive, a bit like in the case of a hibernating animal, whose agency is temporarily "on hold". 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 certainly think of such a system as capable of taking initiatives. Similarly, robotic pets may be explicitly designed to be spontaneous and to take initiatives, while pursuing longer-term goals, such as companionship with a human being. Arguably, 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.

[7] 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).

5.5. 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, while the latter is partly 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. Basically, 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 exclusively human, but could be present in other complex animals too. By contrast, the capacity to bear moral responsibility requires a richer form of intentional 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 so AI systems with free will may plausibly 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 underlying 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 the explanatory logic of choice-making agency is indispensable, the system may be said to have free will in a practically relevant, non-mysterious sense. The system will then satisfy a necessary (albeit not by itself sufficient) condition for bearing moral responsibility. Anyone who considers it desirable for AI systems to function as *moral* agents should find this conclusion congenial.

References

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

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.

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. (1987). *The Intentional Stance.* Cambridge, MA: MIT 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: 349–379.

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

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.) (2011). *The Oxford Handbook of Free Will*. 2nd Edition. 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. (2019). *Why Free Will is Real*. Cambridge, MA: Harvard University Press.

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. (2024). “Decision theory presupposes free will.” https://philpapers.org/archive/LISDTP.pdf

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

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.

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

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.

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.

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.

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.