philosophies

MDPI

Article

How Can a Machine That Is Conscious and Chooses Freely Be Built?

Abraham Meidan

WizSoft, Syosset, New York, NY 11791, USA; abraham@wizsoft.com

Abstract

This paper does not attempt to resolve the metaphysical question of whether consciousness and free will *really exist*, nor whether machines could literally *possess* them. Instead, it pursues a more modest and, I suggest, more fruitful aim: to show that it is possible, in principle, to construct machines that *think they are conscious and think they choose freely*, in essentially the same way that human beings do. To address this question, I identify functional requirements for systems that think they are conscious and think they choose freely. These include subjective self-report, opacity of underlying mechanisms, semantic competence, memory-based self-models, deliberation among alternatives, counterfactual reasoning, and practical unpredictability. I then sketched preliminary computational architectures showing how these capacities could, in principle, be realized using existing or foreseeable technologies. No single component is novel; what matters is their integration into a unified, self-referential system.

Keywords: consciousness; free will; machines; qualia; meaning; counterfactuals

1. Introduction

Throughout the history of philosophy—and with increasing intensity in recent decades—scholars across multiple disciplines have grappled with a set of foundational questions: What is consciousness? What is free choice, or free will? How can human consciousness and free choice be explained? Do non-human animals possess consciousness or the capacity for free choice? And, most recently, can artificial systems such as computers or machines be conscious and choose freely?

These questions have been addressed from a wide range of perspectives, including philosophy, psychology, physiology, neuroscience, and artificial intelligence, yielding a vast and often fragmented literature. Classical philosophical treatments range from Cartesian dualism to empiricist and materialist accounts, while contemporary discussions increasingly draw on cognitive science, neuroscience, and computational modeling. Despite this diversity, no consensus has emerged regarding either the nature of consciousness or the conditions under which free choice is possible.

In the present paper, I focus on a more specific and constructive question: *Is it possible to build a machine that is conscious and chooses freely—and, if so, how?* Rather than asking whether machines *could in principle* possess consciousness or free will in some abstract metaphysical sense, I examine what would be required, in practice, to design and construct such a system. This question lies at the intersection of artificial intelligence and the philosophy of mind and bears directly on ongoing debates concerning strong AI, machine autonomy, and the relationship between physical processes and mental phenomena.

Check for updates

Academic Editor: Marcin J. Schroeder

Received: 1 February 2026
Revised: 8 May 2026
Accepted: 8 May 2026
Published: 13 May 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

If such a machine can indeed be built, it may also provide a new basis for explaining consciousness and free choice in human beings. Historically, scientific explanations of physiological and psychological phenomena have often relied on analogies with machines and artifacts available at the time. These analogies have served not merely as metaphors, but as heuristic devices guiding hypothesis formation and experimental inquiry.

A well-known example is William Harvey's discovery of the circulation of the blood and the function of the heart. Harvey's account was informed by the analogy of a pump: the heart was understood as a device that propels blood through a closed system, in a manner analogous to mechanical pumps circulating water. This mechanistic analogy did not reduce the heart to an artifact, but rather helped to illuminate its functional organization and causal role within the organism. Similar explanatory strategies have since been employed throughout physiology, from respiration to neural signaling.

By the same reasoning, the successful construction of a machine that is conscious and capable of free choice would provide support—though not proof—for the hypothesis that human consciousness and free choice arise from organized physical processes of a broadly similar kind. The claim is not that human beings are machines in a simplistic sense, but that insight into artificial systems exhibiting consciousness and agency may help to demystify these phenomena and render them amenable to scientific explanation. In this respect, artificial models may play a role analogous to that of mechanical models in early modern physiology: they function as concrete realizations of theoretical possibilities and as tools for exploring the conditions under which complex capacities can emerge.

2. Some Remarks on the Formulation of the Question

The question is often posed as follows: *"Is it possible to build a machine that has consciousness and free will?"* This formulation, however, implicitly treats consciousness and free will as if they were *entities*—that is, as discrete things that a system might possess or lack. This assumption is problematic and obscures the nature of the inquiry.

If consciousness and free will are understood as *physical entities*, no such entities have been identified in the human brain or elsewhere in the body, despite extensive investigation in neuroscience and cognitive science. Contemporary research has uncovered complex neural processes and functional organizations correlated with conscious experience and decision-making, but it has not revealed any identifiable object or substance corresponding to consciousness or free will as such (Crick, [1]; Churchland, [2]).

If, on the other hand, consciousness and free will are conceived as *non-physical entities*, this assumption conflicts with contemporary science, which generally seeks to explain phenomena in terms of physical processes, functional organization, and causal mechanisms. While non-physicalist views continue to be defended in philosophy, they tend to resist integration into empirical research programs (Kim, [3]; Papineau, [4]).

A more precise and methodologically conservative approach is therefore to treat consciousness and free will not as entities, but as *properties* or *capacities* of systems. On this view, the relevant question is not whether a machine can contain or instantiate some mysterious entity called "consciousness" or "free will," but whether it can exhibit the kinds of properties and behaviors that, in humans, are associated with being conscious and choosing freely. The question should thus be reformulated as: *"Is it possible to build machines that are conscious and choose freely?"* This shift allows us to bypass contentious metaphysical debates concerning substances or essences and to focus instead on observable and functional characteristics.

However, even this revised formulation remains problematic. Asking whether a machine *is* conscious or *really* chooses freely risks reintroducing metaphysical disputes in a different guise, since the criteria for ascribing consciousness or free will remain controver-

sial. I therefore propose a further refinement: *"Is it possible to build a machine that thinks it is conscious and chooses freely, in the same sense that humans think they are conscious and choose freely?"*

This reformulation shifts the focus from metaphysical speculation to the psychological and cognitive phenomenon of *self-ascription*. Human beings ordinarily experience themselves as conscious agents who make free choices, regardless of whether such experiences can be given a final metaphysical justification. By concentrating on the mechanisms that generate these self-ascriptions—beliefs, reports, deliberative processes, and patterns of behavior—we obtain a question that is both empirically tractable and theoretically productive. This approach aligns with methodological strategies that prioritize explanation of cognitive capacities and subjective reports over ontological commitments concerning the ultimate nature of mind and agency (Dennett, [5]; Metzinger, [6]).

3. The Problem of Free Choice

In brief, the problem of free choice can be stated as follows. Consider an ordinary action, such as raising one's hand. This event occurs in the physical world and, as such, is presumably governed by natural laws, like any other physical phenomenon. Nevertheless, I have a vivid sense of being free to decide whether or not to raise my hand. If I were to decide not to raise it, the action would not occur. The problem of free choice consists of reconciling this subjective sense of freedom with the apparent determinacy of physical processes.

Philosophers have long associated this problem with the broader dispute between determinism and indeterminism. Determinism holds that every event in the physical world is fully determined by prior states in accordance with natural laws, whereas indeterminism maintains that at least some events are not predetermined. It is often said, for example, that classical Newtonian physics exemplifies determinism, while quantum mechanics introduces genuine indeterminism. According to a widely held view, determinism excludes the possibility of free choice, whereas indeterminism leaves room for it.

However, the claim that indeterminism is compatible with free choice is deeply problematic. To say that an event is not predetermined is to say that it is, at least in part, accidental. To say that an event is freely chosen, by contrast, is not to say that it is accidental, capricious, or random. Thus, while determinism is often thought to preclude free choice, indeterminism does not explain it either, nor does it account for our pervasive sense of agency and control. At present, no widely accepted theory explains how free choice can be coherently integrated with the laws of physics, whether those laws are deterministic or indeterministic (Kane, [7]; Mele, [8]).

It is important to note that the question *"Do we have free choice?"* is logically independent of the question *"Do we have a non-material soul?"*. Suppose that human beings possess a non-material soul. Its influence on bodily actions—such as raising one's hand—would still have to be either deterministic or indeterministic. As argued above, neither option suffices to ground the experience of free choice. Invoking a non-material substance therefore does not resolve the problem, but merely displaces it.

The problem, then, is not whether we have free will—almost everyone concedes that we do, at least phenomenologically—nor is it whether human behavior is governed by physical laws, which is also widely accepted. The difficulty lies in the fact that these two seemingly plausible assumptions appear to be mutually inconsistent (Agassi & Meidan, [9]).

When confronted with a paradox of this kind—where two claims each seem true yet jointly incompatible—we have two general options. Either one (or both) of the claims must be false, or the inconsistency must be shown to be only apparent. It is misguided to argue, as some have done, that physics must be rejected because we possess free will (Nagel, [10]), or conversely, that free will must be dismissed as an illusion because physics leaves no

room for it (Dennett, [11]). The appropriate task is instead to identify a deeper flaw in our conceptual framework and to reformulate the problem in a way that dissolves the apparent contradiction. Such a resolution would not eliminate either claim, but would transform our understanding of both.

For this reason, discussions of free will are more fruitful when they focus on the question of *why we feel that we possess free will*, rather than on the metaphysical question of whether we "really" have it. Similarly, in the context of artificial systems, the relevant question is not whether machines can literally possess free will, but whether we can construct machines that think of themselves as choosing freely, in the same way that humans do.

A parallel argument applies to consciousness. On the one hand, I have the strong intuition that someone observing my brain cannot see what I see—for example, the blueness of a blue patch. On the other hand, the hypothesis that I possess a non-physical soul capable of sustaining private experiences is incompatible with contemporary science. Here too, it is therefore more productive to ask why we think we have private experiences, rather than to postulate entities that lie beyond scientific explanation. Accordingly, when considering conscious machines, the relevant question becomes: *Can we build a machine that thinks it is conscious in the same way that humans think they are?*

In what follows, I defend a positive answer to both of these questions. I begin by outlining a proposal for constructing a conscious machine, and then turn to the possibility of building a machine that thinks of itself as choosing freely.

Some philosophers—most prominently John Searle—have argued that only biological systems can give rise to consciousness and free will (Searle, [12]). I do not address these arguments directly here. However, it is clear that a successful demonstration of a conscious material machine would constitute a powerful refutation of the claim that consciousness necessarily depends on a biological substrate.

4. Requirements for a Machine That Thinks It Is Conscious

What requirements must a machine satisfy in order to *think* that it is conscious in the same way that humans think they are conscious? Drawing on the philosophical and cognitive-scientific literature, I propose the following criteria. These are not intended as necessary and sufficient conditions for consciousness *as such*, but as conditions for the self-ascription of consciousness that characterizes human experience.

1. Subjective experience (qualia)

The machine must be able to report that it has perceptual experiences analogous to human qualia, such as the experience of seeing a blue patch. As is well known, there is no intersubjective method for verifying whether the "blue" I experience is identical to the "blue" experienced by another person; such experiences are essentially first-personal (Thomas Nagel, [13]; Frank Jackson, [14]). Qualia states are supposed to have four features: they should be ineffable, intrinsic, private, and directly apprehensible (Dennett, [15]).

A machine that thinks it is conscious must therefore be capable of making comparable first-person reports—e.g., that it experiences colors, sounds, or sensations—without any guarantee that these reports correspond to anything beyond its own internal states. What matters here is not metaphysical certainty about qualia, but the capacity for subjective self-ascription of experience.

2. Transparency of processes

The machine must undergo experiential processes without direct awareness of the physical or computational mechanisms that generate them. When I see an object, I am aware of the object and its properties, not of the neural processes that give rise to visual

experience. This "transparency" of experience has been emphasized in discussions of consciousness and phenomenal content (Block, [16]; Chalmers, [17]).

A machine that thinks it is conscious must similarly distinguish between *what* it experiences and *how* those experiences are implemented. Its internal architecture should therefore shield low-level causal mechanisms from higher-level experiential representations.

3. Self-awareness (the sense of "I")

The machine must represent itself as a persisting individual capable of making first-person, self-referential judgments such as "I am seeing blue" or "I am thinking." This requirement corresponds to what is often called the *minimal self* or *phenomenal self-model* (Gallagher, [18]; Metzinger, [19]).

Importantly, this does not require a metaphysically substantial self. It requires only an internally coherent self-representation that plays a functional role in perception, memory, and action.

4. Representation of reality

At least some of the machine's conscious experiences must be treated as representations of an external world, even though others may be non-veridical, as in imagination, dreaming, or hallucination. Human consciousness routinely exhibits this dual capacity: it can both track reality and depart from it (Revonsuo, [20]).

A system that cannot distinguish, at least implicitly, between experiences that purport to represent reality and those that do not would lack a central feature of human-like consciousness.

5. Understanding of meaning

The machine must be able to grasp the meaning of at least some of the concepts and sentences it uses, rather than merely manipulate symbols according to formal rules. This requirement addresses a central concern in the philosophy of artificial intelligence, most famously articulated in the Chinese Room argument (Searle, [12]).

For present purposes, "understanding" need not be construed in a strong, irreducible sense. It suffices that the machine's internal representations are semantically grounded in its interactions with the world and with itself, rather than being purely syntactic (Harnad, [21]; Dennett, [5]).

6. Memory

Finally, the machine must possess memory, both to integrate experiences over time and to sustain a coherent sense of self and agency. Episodic and autobiographical memory play a crucial role in human consciousness, enabling continuity of experience and the construction of personal identity (Tulving, [22]).

Without memory, a system might generate isolated experiential reports, but it would lack the temporal unity characteristic of conscious self-ascription.

5. A Preliminary Proposal for Building a Machine That Thinks It Is Conscious

Following the general direction suggested by Daniel Dennett [15], I now outline a preliminary proposal for designing a machine (computer) that *thinks* it is conscious, in a manner analogous to the way human beings think they are conscious. What follows is not a complete blueprint but a first sketch, intended to support the more modest claim that the construction of such machines is, in principle, possible.

Consider a machine endowed with the following capabilities.

1. Basic logical, mathematical, and probabilistic competence

The machine is equipped with software implementing a core set of principles from logic, mathematics, and probability theory. It is not necessary to encode the full body of human knowledge in these domains. Rather, a minimal set of deductive and probabilistic

rules suffices to enable elementary reasoning, comparison of hypotheses, and evaluation of uncertainty (Newell & Simon, [23]; Pearl, [24]).

2. Logical inference

The machine is capable of generating logical consequences from the propositions stored in its memory. To avoid combinatorial explosion, inference is constrained by limits on derivation depth, computational resources, or relevance criteria. Such bounded rationality reflects both practical engineering considerations and well-known limits of human reasoning (Simon, [25]).

3. State monitoring

A dedicated monitoring module continuously records which propositions are stored in memory at each moment and preserves a history of these states. This allows the system to represent aspects of its own internal activity and to generate meta-level descriptions such as "I am currently processing visual information" or "I am uncertain about this conclusion." Such self-monitoring mechanisms have long been regarded as central to higher-order theories of consciousness (Rosenthal, [26]).

4. Perceptual input and formal representation

The machine is connected to a range of sensory input devices—such as cameras, microphones, pressure sensors, temperature sensors, and chemical detectors. A translation module converts incoming signals into structured symbolic or sub-symbolic representations, such as predicate-calculus sentences or feature-based representations, which are then stored in memory.

Contemporary systems already perform parts of this task. For example, vision-based applications such as Google's Lookout translate visual input into structured descriptions for downstream processing. While such systems fall far short of human perception, they illustrate the feasibility of automated perceptual encoding.

5. Associative mechanisms

The machine is able to generate associations among stored propositions, analogous to human associative processes. These associations may be based on shared concepts, temporal contiguity, spatial proximity, or statistical co-occurrence. Associative architectures of this kind play a central role in both classical connectionism and contemporary machine-learning systems (Rumelhart & McClelland, [27]; Anderson, [28]).

6. Hypothesis generation

A further module generates generalizations about the world, such as "all ravens are black," and stores them in memory as tentative hypotheses. One possible implementation follows the approach proposed by Agassi and Meidan [29]. According to this approach, the system:

1. Evaluates the *degree of strangeness* of stored propositions. In principle, any property might occur randomly with any other; thus, the discovery of stable, non-random patterns increases the perceived strangeness of the world.
2. Generates hypotheses that reduce this strangeness by:
    - Formulating generalizations that explain surprising regularities.
    - Retrieving associated propositions from memory to test whether they can account for the anomaly.
3. When multiple mutually inconsistent explanations are available, adopt the simplest one, in accordance with the principle of parsimony.

This strategy echoes long-standing ideas about conjecture formation and explanatory power in the philosophy of science (Popper, [30]; Lipton, [31]).

7. Memory management

Finally, the machine includes a memory-management mechanism that prevents uncontrolled growth of stored information. One simple strategy is FIFO (First In, First Out) deletion, though more sophisticated relevance- or utility-based forgetting mechanisms could be employed. Such constraints approximate the dynamic balance between retention and forgetting characteristic of human memory and cognition (Anderson & Schooler, [32]).

All of the above capabilities can, in principle, be implemented using existing computational technologies. Indeed, most of them—or close analogues—are already realized in contemporary software systems. What is novel here is not any single component, but their integration into a unified architecture capable of generating self-referential representations and sustained self-ascriptions of consciousness.

6. Meeting the Requirements for a Machine That Thinks It Is Conscious

The proposed architecture satisfies the requirements outlined above for a system that *thinks it is conscious*, in a manner analogous to how human beings think they are conscious. The point is not that the machine literally possesses consciousness in some metaphysical sense, but that it exhibits the same structural features that underwrite human self-ascriptions of consciousness.

First, the machine lacks access to the details of its own physical implementation. It does not "know" how its hardware operates, just as human beings are not aware of the neuronal mechanisms underlying their mental processes. The machine's awareness is restricted to the *temporal* unfolding of its internal states and operations, not to their spatial realization in hardware. In this respect, it resembles human cognition, where experience presents itself as a sequence of thoughts, perceptions, and decisions without revealing their physical basis. This asymmetry between experiential awareness and physical mechanism was already emphasized by René Descartes, and remains a central theme in contemporary discussions of consciousness (Block, [16]; Chalmers, [17]).

Some philosophers claim that the brain represents its own states in a way that makes them seem to have qualitative properties, but they do not actually have these properties (Frankish [33], Kammerer [34]). The idea presented in this paper is similar: People think they have qualia states, but they are probably wrong about the features they assign to these states. And we can build machines that behave in the same way.

Second, the machine is capable of understanding meaning, provided that understanding a proposition P is construed as the ability to draw appropriate inferences from P in conjunction with other stored propositions. This inferential conception of meaning is characteristic of the tradition initiated by Wilfrid Sellars [35] and further developed by Robert Brandom [36]. Given that the machine acquires empirical information from sensory input (capability #4), generates generalizations (capability #6), and applies logical and probabilistic rules (capabilities #1 and #2), it can derive the inferential consequences that constitute semantic competence on this account.

Third, the machine assigns meanings to perceptual categories such as the color blue. It recognizes, for example, that "blue" is a color and that it resembles perceptual inputs encountered in situations S_1 and S_2 (capability #4). However, this semantic competence does not suffice to determine whether the "blue" experienced by Machine A is identical to the "blue" experienced by Machine B. As with human beings, there is no privileged standpoint from which such comparisons could be made. In this respect, the machine would display something structurally analogous to qualia: private, first-personal aspects of experience that resist intersubjective comparison (Nagel, [13]; Jackson, [14]).

Fourth, the machine would regard itself as a single, persisting individual in virtue of its memory. By continuously monitoring and recording its internal states (capability #3)

and integrating them over time, the system constructs a coherent self-model that links past, present, and anticipated future states. This mirrors influential accounts of human selfhood that emphasize memory and temporal continuity rather than the existence of a substantial ego (Tulving, [22]; Metzinger, [19]).

Fifth, some philosophers claim that knowledge of the external world, as well as self-knowledge, are central mechanisms by which conscious systems maintain their identity and survival. The system suggested here meets this requirement by applying logical inference (capability #2), perceptual input (capability #4), and hypothesis generation (capability #6).

Taken together, these features suffice to explain why such a machine would *think of itself* as conscious. It would report experiences, distinguish appearance from mechanism, employ meaningful concepts, and represent itself as a temporally extended subject of experience. From the machine's own perspective, there would be no principled difference between its situation and that of a human being reflecting on their own consciousness.

7. Requirements for a Machine That Thinks It Chooses Freely

I turn now to the problem of free choice. As noted at the beginning of this paper, the question is not *"Do we have free choice?"* but rather: *Why do we feel that we have free choice, and how can this feeling be explained in a way that is consistent with accepted science?*

A natural but misleading idea is that I have free choice because another person—or even an omniscient being—cannot accurately predict what I am going to choose. This interpretation, however, immediately encounters the difficulties associated with determinism and indeterminism discussed earlier.

If determinism is true, then in principle an agent located outside the system and possessing complete information could predict my next action. (The predictor must be external to the system in order to avoid trivial counterexamples—for instance, a chess opponent who predicts my next move, and thereby induces me to choose differently.) If, on the other hand, indeterminism is true, then some of my choices may indeed be unpredictable because they are partially accidental. But accidental behavior is not consistent with our experience of choosing freely; randomness does not amount to agency.

Thus, unpredictability—whether due to ignorance or indeterminism—cannot by itself explain the sense of free choice. The problem is to explain the phenomenology and structure of agency. The relevant question becomes: *What requirements must a machine satisfy in order to think that it chooses freely in the same way that humans think they do?*

Philosophers working in the compatibilist and naturalistic traditions have proposed a number of criteria. The following list summarizes the most widely accepted ones.

1. Conscious deliberation

Free choice requires that the agent be conscious of its deliberation and of the outcome of that deliberation. Actions performed unconsciously, reflexively, or under hypnosis are typically excluded from the domain of free choice. This requirement links free will directly to conscious awareness and reportability (Dennett, [11]; Rosenthal, [26]).

2. The could-have-done-otherwise condition

An agent is commonly said to act freely only if it *could have acted otherwise* had it chosen to do so. In contemporary compatibilist accounts, this condition is often interpreted counterfactually: if the agent had had different reasons, desires, or intentions, it would have acted differently (Ayer, [37]; Lewis, [38]).

3. Sourcehood (being the origin of one's action)

Free choice requires that the agent be the appropriate source of its action. Actions that result from coercion, manipulation, or external control are not regarded as free, even if they are internally consistent with the agent's beliefs and desires. What matters is that the causal

chain leading to action runs through the agent's own deliberative processes (Frankfurt, [39]; Fischer & Ravizza, [40]).

4. Reasons-responsiveness

A widely accepted condition for free choice is that the agent acts for reasons and is capable of responding appropriately to reasons. An agent who is completely insensitive to reasons—who would act the same way regardless of considerations for or against an action—lacks free will in the relevant sense. Reasons-responsiveness does not require perfect rationality, only a stable capacity to modify behavior in light of reasons (Fischer & Ravizza, [40]; Mele, [41]).

5. Control and self-regulation

Free choice involves a form of control over one's behavior, including planning, self-reflection prior to action, and the psychological capacity to regulate impulses. This notion of control is compatible with deterministic causation and is typically understood in terms of higher-order cognitive capacities rather than metaphysical freedom (Dennett, [11]; Baumeister, [42]).

6. Practical unpredictability by other agents

Although unpredictability is not a metaphysical requirement for free will, it plays an important *phenomenological* role. From the agent's own perspective—and from the perspective of other agents with limited information—freely chosen actions are not reliably predictable. This practical unpredictability arises from the complexity of deliberation, sensitivity to reasons, and self-referential reasoning, rather than from indeterminism (Popper, [43]; Dennett, [44]).

Taken together, these criteria characterize what it is for an agent to *think of itself* as choosing freely. None of them requires indeterministic physics or a non-material soul. Instead, they point toward an account of free choice grounded in conscious deliberation, reasons-responsiveness, self-control, and self-ascription of agency. In the next section, I shall argue that these requirements can, in principle, be satisfied by an appropriately designed artificial system.

8. A Preliminary Proposal for Building a Machine That Thinks It Chooses Freely

I now outline a preliminary proposal for designing a machine (computer) that *thinks it chooses freely*, in a manner analogous to how human beings think they possess free choice. As emphasized earlier, this proposal is not intended as a complete or final model, but as an initial sketch meant to support the claim that constructing such machines is, in principle, possible.

Consider the conscious machine described in the previous sections, augmented with the following additional capabilities. Three further features are required for the self-ascription of free choice.

The additional features required for free choice

1. **Option generation and selection**

The machine must be able to represent multiple possible courses of action and to select one of them on the basis of deliberation.

2. **Opacity of decision-making processes**

The machine must not have direct introspective access to the full causal chain by which its decisions are produced.

3. **Counterfactual reasoning**

The machine must be able to consider counterfactual alternatives—what it *could have done otherwise*—and to represent these alternatives as genuine possibilities.

The first two features are relatively straightforward.

Feature (1) is already realized in many existing computer systems. Programs that play chess, for example, routinely generate multiple candidate moves, evaluate them according to various criteria, and select one as optimal. Such systems already satisfy the minimal requirement of option generation and selection (Newell & Simon, [23]).

Feature (2) can be implemented in the same way as in the case of consciousness, discussed above. Just as the machine lacks access to the low-level physical mechanisms underlying its conscious states, it may also lack direct access to the detailed causal processes that generate its decisions. From the system's own perspective, decisions simply "occur" as outcomes of deliberation, without transparency regarding their physical realization. This opacity mirrors the human experience of agency (Dennett, [11]).

The third feature—counterfactual reasoning—is the most philosophically and technically interesting. It can be implemented by drawing on *possible-worlds semantics*, as developed by David Lewis [45] and Robert Stalnaker [46].

According to possible-worlds semantics, a counterfactual statement refers to what happens in a possible world that is similar to the actual world in relevant respects. Applied to free choice, the claim that an agent *could have done otherwise* can be analyzed as follows:

An agent could have done otherwise if and only if there exists a sufficiently nearby possible world in which:

- The laws of nature are the same as in the actual world;
- The agent's initial conditions are the same as in the actual world, with the exception of the agent's choice as defined in the counterfactual condition;
- Following these conditions, the agent performs a different action.

To implement this idea computationally, the machine must store:

1. A representation of relevant laws or regularities governing its environment and its own operation.
2. A set of factual descriptions of the current situation.

Given a counterfactual condition (for example, "If I had chosen action B instead of A. . ."), the program proceeds by identifying which stored facts are inconsistent with that condition and temporarily ignoring them. It then evaluates whether the remaining facts, together with the relevant laws, entail the counterfactual outcome. This procedure parallels standard formal treatments of counterfactuals in logic and philosophy (Lewis, [45]; Pearl, [47]).

Equipped with this capacity, the machine can represent alternative possible actions as genuine options that it *could have chosen*, even though its actual choice is the result of deterministic or probabilistic processes. From the machine's own perspective, this suffices to ground the sense that it could have done otherwise.

A natural question arises at this point: Is such a conception of free choice compatible with our practice of holding agents morally responsible—for example, justifying punishment for immoral actions freely chosen, while excusing actions performed under coercion?

The answer is affirmative. In everyday moral practice, we already assume that actions are shaped—often deterministically—by an agent's past experiences, character, and expectations. Punishment is justified not because the action was uncaused, but because it influences future deliberation. We believe that when an agent next considers performing an immoral act, the expectation of punishment will enter into its deliberative process and may prevent the choice of that act. This forward-looking, consequentialist rationale for responsibility applies equally to human beings and to suitably designed machines (Dennett, [11]; Fischer & Ravizza, [40]).

Thus, a machine that deliberates, evaluates alternatives, reasons counterfactually, and regulates its behavior in light of anticipated consequences would have all the resources required to *think of itself* as choosing freely—and to be treated, in relevant respects, as an agent.

9. Conclusions: What Has—And Has Not—Been Shown

The central aim of this paper has not been to resolve the metaphysical questions of whether consciousness and free will *exist* in some ultimate sense, nor to determine whether machines can literally *possess* them. Rather, the guiding thesis has been explanatory: *to show that it is possible, in principle, to construct machines that think they are conscious and think they choose freely, in essentially the same way that human beings do*.

The paper has advanced this thesis by reframing traditional philosophical problems. Instead of asking whether consciousness and free will are entities, substances, or irreducible properties, I have treated them as *self-ascriptions* grounded in cognitive architecture. The key explanatory task, on this view, is not to explain how consciousness or free will transcend physical law, but to explain why agents—human or artificial—experience themselves as conscious and free within a law-governed world.

To this end, I proposed a set of functional and architectural requirements for systems that think they are conscious and think they choose freely. These requirements include subjective self-report, opacity of underlying mechanisms, semantic competence, memory-based self-models, deliberation among alternatives, counterfactual reasoning, and practical unpredictability. I then outlined preliminary computational architectures showing how these requirements could, in principle, be realized using existing or foreseeable technologies.

The proposal advanced here should be clearly distinguished from prevailing approaches based on large language models (LLMs). The central difference lies in the level of explanatory transparency. The present proposal is designed to provide a fully articulated account of the mechanisms generating the machine's outputs, whereas contemporary LLMs are often characterized as opaque or "black-box" systems, in the sense that their internal representations and decision processes are not readily accessible.

In addition, the present account diverges from some speculative interpretations of LLMs according to which sufficiently advanced systems might give rise to genuine consciousness. The claim defended here is more modest and more precise. It does not posit that consciousness will emerge in such a system. Rather, it maintains that a suitably designed machine can be constructed that will report or assert that it is conscious and choose freely, in a manner structurally analogous to human reports of consciousness and free choice. This position is compatible with philosophical views according to which such reports do not require the existence of irreducible phenomenal properties, but may instead arise from underlying cognitive and representational mechanisms.

The ideas developed in this paper are broadly compatible with the proposals advanced by Rao Mikkilineni and William P. Kelly concerning "Mindful Machines", which operate through a continuous Discover–Reflect–Apply–Share cycle (Mikkilineni & Kelly, [48]), as well as with their proposal to extend the Shannon–Turing model of computation by incorporating a "Governance of Commitments" layer (Mikkilineni & Kelly, [49]). Nevertheless, a detailed examination of these frameworks and their implications lies beyond the scope of the present paper.

What has been shown is the following:

1. **The sense of consciousness and free choice does not require indeterministic physics, non-material souls, or biologically unique substrates.**

Both can be accounted for in terms of self-monitoring, representational opacity, inferential competence, and counterfactual reasoning.

2. **A machine that satisfies the proposed criteria would be structurally indistinguishable, from its own perspective, from a human agent reflecting on its consciousness and freedom.**

Such a machine would report experiences, regard itself as a persisting self, deliberate among alternatives, and believe that it could have acted otherwise.

3. **The traditional paradoxes surrounding free will—determinism versus indeterminism—lose much of their force once the focus shifts from metaphysical freedom to the phenomenology and function of agency.**

Neither randomness nor metaphysical indeterminacy is required to ground the experience of choice.

4. **The proposed account is compatible with practices of responsibility, learning, and moral evaluation.**

Both human and artificial agents can be influenced by expectations, sanctions, and future-oriented reasoning, which suffices for many practical purposes associated with responsibility.

Equally important, however, is what has *not* been shown.

Firstly, this paper does **not** demonstrate that machines built along these lines would be conscious or free in any metaphysically robust sense. Those questions remain open, and may ultimately be ill-posed. The paper proposes a non-metaphysical account of consciousness without rejecting metaphysical approaches. Its aim is not to refute metaphysical theories, but to introduce an additional explanatory option that operates independently of them. Secondly, the paper does **not** claim that current artificial systems already meet these requirements, nor that implementing them would be trivial. The proposal is architectural and conceptual, not a report of an existing implementation. Thirdly, nothing here establishes that human consciousness or free will *reduces* to the proposed mechanisms; the argument is abductive and explanatory, not reductive or eliminative.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The author declares no conflicts of interest. The author certifies that he has no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

1. Crick, F. *Astonishing Hypothesis*; Scribner: New York, NY, USA, 1994.
2. Churchland, P.M. *Matter and Consciousness*; MIT Press: Cambridge, MA, USA, 1988.
3. Kim, J. *Mind in a Physical World*; MIT Press: Cambridge, MA, USA, 1998.
4. Papineau, D. *Thinking About Consciousness*; Oxford University Press: Oxford, UK, 2002.
5. Dennett, D.C. *The Intentional Stance*; MIT Press: Cambridge, MA, USA, 1989.
6. Metzinger, T. *Being No One*; MIT Press: Cambridge, MA, USA, 2003.
7. Kane, R. *The Significance of Free Will*; Oxford University Press: Oxford, UK, 1996.
8. Mele, A. *Free Will and Luck*; Oxford University Press: Oxford, UK, 2001.
9. Agassi, J.; Meidan, A. *Beg to Differ*; Springer: Berlin/Heidelberg, Germany, 2016.
10. Nagel, T. *The View from Nowhere*; Oxford University Press: Oxford, UK, 1986.
11. Dennett, D.C. *Elbow Room*; Bradford Books: New Haven, CT, USA, 2015.
12. Searle, J.R. Minds, Brains, and Programs. *Behav. Brain Sci.* **1980**, *3*, 417–457. [CrossRef]
13. Nagel, T. What Is It Like to Be a Bat? *Philos. Rev.* **1974**, *83*, 435–450. [CrossRef]
14. Jackson, F. Epiphenomenal Qualia. *Philos. Q.* **1982**, *32*, 127–136. [CrossRef]

15. Dennett, D.C. *Consciousness Explained*; Penguin: New York, NY, USA, 1993.
16. Block, N. On a Confusion about a Function of Consciousness. *Behav. Brain Sci.* **1995**, *18*, 227–287. [CrossRef]
17. Chalmers, D.J. *The Conscious Mind*; Oxford University Press: Oxford, UK, 1996.
18. Gallagher, S. Philosophical Conceptions of the Self. *Trends Cogn. Sci.* **2000**, *4*, 14–21. [CrossRef] [PubMed]
19. Metzinger, T. *The Ego Tunnel*; Basic Books: New York, NY, USA, 2009.
20. Revonsuo, A. Consciousness, Dreams and Virtual Realities. *Philos. Psychol.* **1995**, *8*, 35–58. [CrossRef]
21. Harnad, S. The Symbol Grounding Problem. *Phys. D* **1990**, *42*, 335–346. [CrossRef]
22. Tulving, E. Episodic Memory: From Mind to Brain. *Annu. Rev. Psychol.* **2002**, *53*, 1–25. [CrossRef] [PubMed]
23. Newell, A.; Simon, H.A. Computer Science as Empirical Inquiry. *Commun. ACM* **1976**, *19*, 113–126. [CrossRef]
24. Pearl, J. *Probabilistic Reasoning in Intelligent Systems*; Morgan Kaufmann: San Francisco, CA, USA, 1988.
25. Simon, H.A. *Models of Man*; Wiley: Hoboken, NJ, USA, 1957.
26. Rosenthal, D.M. *Consciousness and Mind*; Oxford University Press: Oxford, UK, 2005.
27. Rumelhart, D.E.; McClelland, J.L. *Parallel Distributed Processing*; MIT Press: Cambridge, MA, USA, 1986.
28. Anderson, J.R. *Rules of the Mind*; Lawrence Erlbaum: Mahwah, NJ, USA, 1993.
29. Agassi, J.; Meidan, A. *Philosophy from a Skeptical Perspective*; Cambridge University Press: Cambridge, UK, 2008.
30. Popper, K.R. *The Logic of Scientific Discovery*; Hutchinson: London, UK, 1959.
31. Lipton, P. *Inference to the Best Explanation*; Routledge: Abingdon, UK, 2004.
32. Anderson, J.R.; Schooler, L.J. Reflections of the Environment in Memory. *Psychol. Sci.* **1991**, *2*, 396–408. [CrossRef]
33. Frankish, K. Illusionism as a Theory of Consciousness. *J. Conscious. Stud.* **2016**, *23*, 11–39.
34. Kammerer, F. The illusion of conscious experience. *Synthese* **2021**, *198*, 845–866. [CrossRef]
35. Sellars, W. *Empiricism and the Philosophy of Mind*; Harvard University Press: Cambridge, MA, USA, 1997. (Original work published 1956).
36. Brandom, R.B. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*; Harvard University Press: Cambridge, MA, USA, 1994.
37. Ayer, A.J. Freedom and Necessity. In *Philosophical Essays*; Macmillan: New York, NY, USA, 1954.
38. Lewis, D. Are We Free to Break the Laws? *Theoria* **1981**, *47*, 113–121. [CrossRef]
39. Frankfurt, H.G. Alternate Possibilities and Moral Responsibility. *J. Philos.* **1969**, *66*, 829–839. [CrossRef]
40. Fischer, J.M.; Ravizza, M. *Responsibility and Control*; Cambridge University Press: Cambridge, UK, 1998.
41. Mele, A. *Motivation and Agency*; Oxford University Press: Oxford, UK, 2003.
42. Baumeister, R.F. *Free Will in Scientific Psychology*; APS—Association for Psychological Science: Washington, DC, USA, 2008.
43. Popper, K.R. *The Open Universe*; Hutchinson: London, UK, 1982.
44. Dennett, D.C. *Freedom Evolves*; Penguin: New York, NY, USA, 2004.
45. Lewis, D. *Counterfactuals*; Harvard University Press: Cambridge, MA, USA, 1973.
46. Stalnaker, R. A Theory of Conditionals. In *Studies in Logical Theory*; American Philosophical Quarterly Monograph Series; Basil Blackwell: Oxford, UK, 1968.
47. Pearl, J. *Causality: Models, Reasoning, and Inference*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2009.
48. Mikkilineni, R.; Kelly, W.P. From Static Prediction to Mindful Machines: A Paradigm Shift in Distributed AI Systems. *Computers* **2025**, *14*, 541. [CrossRef]
49. Mikkilineni, R.; Kelly, W.P. The Missing Layer in Modern IT: Governance of Commitments, Not Just Compute and Data. *Computers* **2026**, *15*, 275. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.