philosophies

IMPACT FACTOR 0.7

CITESCORE 1.7

Article

Moral Judgment Without Moral Agency: Why Judgment Accuracy Does Not Ground Moral Agency in the Age of AI

Abdurrazak Gültekin

MDPI

https://doi.org/10.3390/philosophies11050186

philosophies

MDPI

Article

Moral Judgment Without Moral Agency: Why Judgment Accuracy Does Not Ground Moral Agency in the Age of AI

Abdurrazak Gültekin

History of Philosophy Division, Department of Philosophy and Religious Sciences, Faculty of Theology, İnönü University, 44280 Malatya, Türkiye; abdurrazak.gultekin@inonu.edu.tr

Abstract

As large language models increasingly match human performance on moral-judgment tasks, two bodies of literature have developed in parallel. Empirical moral psychology treats moral judgment as an output and benchmarks artificial systems against human responses. The philosophy of artificial moral agency denies agency to such systems mainly on the ground of sentience, an empirically inaccessible criterion that has left the debate stalled. This paper argues that human–AI convergence in moral judgment conceals a categorical asymmetry that neither literature has theorized. Drawing on Spinoza's conatus and Damasio's account of homeostasis and affect, it argues that moral agency presupposes affective-homeostatic self-maintenance as its constitutive ground: the organization from which valuation first arises. This condition is necessary but not sufficient, and the claim is constitutive rather than normative, so it does not derive an "ought" from an "is." Sentience is repositioned as the experiential manifestation of this organization rather than the explanatory terminus. Contemporary AI systems, including those trained with human feedback, are therefore moral judges without moral agency. The paper closes with a behavioral research program, including concrete experimental designs, on the conditions under which people attribute moral agency to human and artificial agents.

Keywords: moral agency; moral judgment; affect; homeostasis; self-maintenance; artificial intelligence; dual-process theory; moral responsibility; Spinoza; conatus

1. Introduction

A striking finding has been accumulating across moral psychology and AI research. When large language models rate the wrongness of actions in standard moral scenarios, their ratings correlate closely with those of human participants [1], and such models encode a recognizably human-like sense of what is right and wrong to do [2]. In a blinded comparison, people even rated a model's moral evaluations as superior in quality to human ones on most dimensions [3]. These results have led researchers to ask whether such models might stand in for human participants in moral-psychological research [1]. Recent work in Behavioral Sciences sharpened the question by comparing how people judge the same act when its author is described as a human or as an artificial agent [4]. At the level of the verdict, the distance between human and machine has narrowed to a degree that would have seemed implausible a decade ago.

It is tempting to read this convergence as a claim about competence. If a system reliably produces the judgments a competent moral evaluator would produce, it is easy to credit it with moral competence. From there it is a short step to the thought that it participates, however minimally, in moral agency. This slide is rarely made explicit, but it

Check for updates

Academic Editor: Craig Reeves

Received: 3 September 2026
Revised: 4 October 2026
Accepted: 5 October 2026
Published: 12 October 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.

structures much of the surrounding discourse. The central claim of this paper is that the convergence is real, but the inference is mistaken. Seeing why requires a distinction that the two governing bodies of literature have kept apart.

On one side stands empirical moral psychology. Its dominant framework, the dual-process account, treats a moral judgment as the output of interacting fast-affective and slow-deliberative processes [5–7]. Its experimental apparatus is built to measure and explain that output: trolley and footbridge dilemmas, manipulations of cognitive load and time, comparisons across agent types. When this apparatus is turned on artificial systems, the natural quantity to compare is again the output: does the model judge as humans judge? The framework is powerful, but it is built to study judgments. It therefore tends to treat the judgment as the thing whose presence settles the matter.

On the other side stands the philosophy of artificial moral agency. Here the question is whether machines that behave morally can be moral agents at all. The most influential negative answers turn on consciousness or sentience. On this view, an algorithm cannot be a moral agent because there is nothing it is like to be one, and moral responsibility requires a subject who can experientially grasp what is at stake [8]. This is a serious position. But it locates the decisive difference in phenomenal experience, a property we cannot observe or adjudicate from outside. The debate has accordingly stalled between those who deny machine sentience and those who insist it cannot be ruled out. Relational approaches have tried to bypass the impasse by grounding moral status in social relations rather than in intrinsic properties [9,10]. These approaches shift the question rather than answer it.

The two bodies of literature pass each other. One measures whether machines judge as we do and risks treating convergence as evidence of moral standing. The other denies that standing but anchors the denial in a property no experiment can reach. Neither connects moral agency to the layer that affective neuroscience and a long strand of ethical theory both identify as the origin of valuation: the affective, homeostatic self-maintenance of a living body with a stake in its own persistence [11–13]. The consequences of this omission have not been theorized. Suppose valuation originates in something a system either enacts or fails to enact. Then two systems can converge completely on the output of valuation while differing categorically in whether they value at all. Judgment-convergence would then be fully compatible with the absence of moral agency in the machine. The reason would not be that the machine lacks a hidden inner light. It would be that the machine lacks the self-maintaining ground from which moral judgment, in the human case, borrows its point.

This is the paper's thesis. Moral agency is not reducible to the capacity to produce normatively appropriate moral judgments. It presupposes affective-homeostatic self-maintenance, a conatus-like striving to persist, as its constitutive ground. The term ground is deliberate and narrow. It names neither a metaphysical grounding relation nor a mere causal antecedent. It names a constitutive and explanatory condition: the organization that makes valuation, and therefore morally significant concern, possible at all. Self-maintenance is claimed to be necessary, not sufficient. As Biological Normativity Is Not Yet Moral Normativity Section argues, not every form of valuation is moral, and a further account is owed of how biological normativity becomes the moral kind. The thesis is also constitutive rather than normative. It says what a being must be in order to be a moral agent. It does not say that the good is whatever preserves the agent (see Section 7.2). On this view, a system that reproduces human moral verdicts is a moral judge without moral agency. It produces normatively appropriate evaluations without the affective-homeostatic organization that would make those evaluations its own stakes. It reproduces the content of valuation without instantiating its condition. It reproduces what humans value without being a being for whom anything is valuable.

1.1. Key Terms

The argument turns on distinctions that ordinary usage blurs, so the central terms are fixed here.

Valuation is the most basic of them. A system values when states of itself or its environment register as better or worse relative to norms that the system itself maintains. Valuation in this sense is a relation between a system and its conditions of persistence, not a cognitive act. It must be distinguished from scoring, in which outputs are ranked against a metric supplied from outside, such as a loss function, a reward model, or a rubric (see Section 5.2). Throughout, "valuation" without qualification means endogenous valuation.

Affect is the felt registering of valuation. In organisms like us, the difference between better and worse is not only instantiated but undergone, as feeling [12,14]. Affect therefore presupposes valuation, but not every valuation is affective. A bacterium moving up a nutrient gradient instantiates a better and a worse, yet on most views nothing is felt. Affect in this foundational sense also differs from how the dual-process literature typically uses the word. There, "affect" usually names a fast process that contributes to a verdict. Here it names the medium in which a being's stake appears to that being (Section 2).

Moral judgment is an evaluative verdict with moral content about an action, a character, or a state of affairs. Two senses must be separated. In the output sense, a moral judgment is a verdict token, such as a wrongness rating, which any sufficiently capable system can produce. In the act sense, a moral judgment is a judging performed by an agent who stands behind it and is answerable for it. Empirical work measures the first; responsibility practices presuppose the second. When this paper calls AI systems "moral judges," it uses the output sense only.

Moral agency, finally, is a status rather than an output. It is the standing of a being whose conduct can be attributed to it as its source, so that praise, blame, and responsibility apply. Moral Agency, Moral Standing, and Moral Patienthood Section further distinguishes it from moral standing and patienthood.

These terms are ordered. Valuation is the condition of affect. Affect, once moralized (Biological Normativity Is Not Yet Moral Normativity Section), is the condition of moral concern. Moral judgment in the act sense is one expression of moral agency. Moral judgment in the output sense, by contrast, can occur either at the end of this chain or entirely without it. That possibility is the subject of this paper.

1.2. Outline

Section 2 identifies an assumption in how the dual-process frame is operationalized. Section 3 reconstructs the homeostatic-affective ground of valuation from Spinoza and Damasio, and Biological Normativity Is Not Yet Moral Normativity Section marks the threshold between biological and moral normativity. Section 4 distinguishes judgment from agency, and Moral Agency, Moral Standing, and Moral Patienthood Section separates moral agency from moral standing and patienthood. Section 5 relocates the ground of agency and repositions sentience. Section 5.1 specifies which self-maintaining systems qualify, and Section 5.2 tests the criterion against current AI architectures. Section 6 reinterprets the empirical convergence. Section 7 answers objections, including the is–ought objection. Section 8 converts the thesis into a behavioral research program with concrete designs, and Section 9 concludes. The paper offers no verdict on any particular machine. It relocates the question away from what machines can judge and toward what, in judging, they would have to be.

https://doi.org/10.3390/philosophies11050186

2. The Dual-Process Frame and How Affect Is Operationalized

Most contemporary work on moral cognition descends from the proposal that moral judgments issue from two dissociable kinds of processing. Emotional responses, engaged especially by up-close and personal harms, push toward characteristically deontological verdicts. Controlled cognitive processing, when recruited, supports characteristically consequentialist ones [5,7]. The neuroimaging work that launched this picture read differences in judgment across dilemma types as differences in the balance of affective and deliberative engagement [5]. The intuitionist tradition alongside it argued that moral verdicts arrive first, as quick affect-laden intuitions, and that reasoning follows to justify them [6]. The tradition is internally diverse, and no careful theorist reduces affect to a single function. Its experimental apparatus nonetheless shares a common object: the judgment, understood as a verdict about a case and explained by the processes that produce it.

This framework has been remarkably generative, and its extension to artificial agents is natural. A study in Behavioral Sciences asked how people evaluate identical actions performed by human and AI agents [4]. It worked squarely within the dual-process idiom, using the trolley and footbridge cases because they are thought to recruit the two processes differently. The results bear directly on the present argument. In the impersonal trolley case, the type of agent, not only the action, moved judgment: an AI agent's intervention was rated more objectionable and more blameworthy than the same intervention by a human. In the more visceral footbridge case, the action dominated, and the utilitarian option was judged more wrong regardless of who performed it [4]. Related work finds that people are averse to machines making moral decisions, an aversion mediated by the perception that machines can neither fully think nor feel [15]. Other studies show that people apply different moral norms to robot and human agents facing the same dilemma [16]. People, in short, do not treat a machine's morally identical act as morally identical. Something about the agent enters the evaluation even when the behavior is held fixed.

The frame has also begun to interrogate itself, and this opens the space the present paper occupies. A recent reanalysis and preregistered replication failed to detect an effect of chronic stress on moral decision-making that earlier work had reported. The authors concluded cautiously that the reported effect may be less robust than assumed [17]. More pointedly, a conceptual critique argues that the dual-process account rests on untested assumptions and that its standard methods cannot falsify it. On this critique, the affective/cognitive partition does more work than the evidence licenses [18]. These are not complaints from outside. They are the framework tightening its own screws.

What this internal criticism does not yet question is a commitment carried less by the theory's explicit claims than by its operationalization. The experimental program treats a moral judgment as the joint output of an affective and a cognitive process. It therefore tends to characterize affect by its contribution to the verdict: a fast channel that, in certain dilemmas, delivers the answer before deliberation can. Affect and cognition are measured as two roads to the same destination, and the empirical work largely charts which road is taken under which conditions. Call this the routing assumption: the tendency, in operationalization if not in doctrine, to treat affect's contribution to morality as exhausted by its role in generating verdicts.

The routing assumption is not obviously wrong, but it has a consequence that becomes visible only when artificial systems enter the picture. If affect is one route to the verdict, a system might in principle reach the same verdicts by the other route, or by a functional substitute for both. It would then possess everything morally relevant that the judgment carries. Machine–human convergence would be exactly what it appears to be: two systems reaching one destination by different roads. The alternative developed here is that affect, in its foundational role, is not a route to the verdict at all. It is the condition under which

there is a destination: the ground on which anything shows up as mattering, as worth a verdict in the first place. If so, a system could travel the deliberative road flawlessly and produce every verdict a moral agent would produce, yet still not do what a moral agent does. It would never have possessed the stake that makes a verdict a moral one rather than a classification. To see why affect is ground rather than route, we must leave the psychology of judgment and ask where valuation comes from.

3. The Homeostatic-Affective Ground of Valuation

Long before there are moral judgments, there are things that matter to a creature. The oldest philosophical account of why anything matters begins not with reason but with the effort to persist. Spinoza names this effort. Each thing, as far as it can by its own power, strives to persevere in its being, and this striving, conatus, is nothing other than the thing's actual essence [19] (E3P6–P7). [1] From this premise Spinoza derives a theory of value more radical than it first appears. Good and evil are not properties that things possess independently of any creature and that a creature then perceives. They are relational, indexed to the striving itself. We call good what is useful to preserving our being, that is, what increases or aids our power of acting, and bad what diminishes or restrains it [19] (E4D1–D2, E4P8Dem). Value is not read off the world by a judging intellect. It is generated by a self-maintaining being whose persistence is at stake in its encounters. That being's affects of joy and sadness are the registering of those encounters as passages to greater or lesser perfection [19] (E3P11S).

Two features of this account matter for what follows. First, the power at issue is not bare biological survival. For Spinoza, what the mind strives for insofar as it reasons is understanding, and its greatest good is knowledge of God [19] (E4P26–P28). Second, the account is not confined to the individual. The good that a person who pursues virtue wants for herself, she also desires for others [19] (E4P37). Both points return in Section 7.2.

Deleuze's reading turns this into a method he calls ethology. To understand a being ethically is to ask what it can do and what it can affect and be affected by. Value then follows from the composition and decomposition of relations, not from conformity to an externally imposed norm [20]. The ordering matters. Evaluation is not added by a separate faculty to a value-neutral representation of the world. It is coextensive with the striving of a being that has something to lose. There is no valuing without a valuer whose own persistence supplies the currency in which things register as better or worse.

What was a metaphysical thesis in the seventeenth century has acquired, over the last three decades, a strikingly convergent naturalistic formulation, though the two must not be collapsed. Damasio's work begins from patients whose emotional processing was impaired by focal brain damage. Stripping feeling from reasoning did not yield a cooler, more efficient reasoner. It yielded a person whose personal and social decision-making collapsed [11]. The resulting somatic-marker hypothesis holds that bodily affective signals do not interfere with good practical judgment. They are a necessary substrate for it, marking options with value before and beneath deliberation [11]. In later work Damasio makes a deeper claim explicit: feelings are the mental expression of homeostasis. Homeostasis is the continuous regulation by which a living system keeps itself within the narrow range of states compatible with its survival [12,13]. On this account, value is not a late cognitive achievement layered onto a homeostatic machine. Feelings are the mental experience of the body's homeostatic state: the organism's states of greater or lesser viability are registered as feelings of greater or lesser well-being [13,21].

It is important to be exact about what this establishes. Damasio supplies a naturalistic account of the biological architecture within which valuation and affect arise. He does not, and need not, supply a theory of moral agency. The step from that architecture to

moral agency is philosophical, and the burden of making it falls on this paper, not on the neuroscience.

The enactive tradition makes a parallel claim from within the philosophy of biology. Following Jonas [22], it treats living systems as the paradigm of entities with a concern for their own being, because their identity depends on continuously producing themselves [23]. Adaptivity, the capacity of an organism to regulate itself with respect to its conditions of viability, is what turns mere self-production into sense-making and the generation of values [24]. This tradition likewise treats a primordial affectivity, a sensitivity to what helps or harms, as constitutive of even basic forms of mind rather than as an optional overlay [14]. Theories that treat emotions as themselves evaluative trace the same dependence of value on a being for whom things can go well or badly, at the level of full human emotion [25,26].

The convergence among these accounts is real but structural, not a matter of identity. Spinoza's conatus and Damasio's homeostasis are conceptually distinct accounts of the relation between self-maintenance and valuation. What they share is that both locate the origin of value in the self-maintenance of a being whose persistence is not guaranteed and is therefore at issue for it. The argument does not claim that Spinoza anticipated homeostasis. Nor does it claim that conatus, an ontological notion that covers all finite modes, is the same as a biological regulatory process. It claims only that the two illuminate the same structural fact from different levels of description. On this shared view, affect is the medium in which the stake of a self-maintaining being is registered. It is not a fast heuristic that a slower faculty might replace. It is the appearing of mattering as such.

Biological Normativity Is Not Yet Moral Normativity

An objection now presses, and meeting it is essential. If valuation arises wherever a self-maintaining being has a stake, valuation is extraordinarily widespread. A bacterium moving up a nutrient gradient, a plant orienting to light, and an animal fleeing a predator all occupy states that are better or worse for them. Does the argument then ascribe moral agency to bacteria and plants? It does not, and the reason exposes a distinction the thesis depends on.

Three levels must be kept apart. At the first, biological normativity, a state is simply better or worse for an organism relative to its persistence. At the second, affective valuation, that difference is registered by the organism as felt: undergone rather than merely instantiated. At the third, moral normativity, conduct is evaluated as right or wrong, fair or unjust, obligatory or forbidden, both for oneself and for others.

The homeostatic-affective account developed here concerns the passage from the first level to the second. It identifies the condition under which anything can matter to a system at all. That condition is necessary for moral agency but plainly not sufficient. A being for whom things can go better or worse does not yet hold conduct to standards, weigh reasons, or treat the stakes of others as making claims on it. What converts affective valuation into the moral kind may be called the moralization of stakes. It requires a self-maintaining affective organization developed enough that stakes can become reasons: reasons for action, and reasons in the evaluation of oneself and others.

This paper does not offer a complete theory of that further step. Its aim is narrower and, for that reason, more secure. It establishes that affective-homeostatic self-maintenance is the constitutive ground without which the higher capacities have nothing to be about. A system lacking this ground cannot possess moral agency, however well it reproduces moral verdicts. Bacteria are excluded not merely because they lack judgment, which they do. They are excluded because biological normativity, without a developed affective organization in which stakes become reasons, does not amount to moral agency. The claim about artificial

systems is stronger still. As Sections 5.1 and 5.2 argue, present systems lack even the first, generative level.

4. Judgment Is Not Agency

The empirical literature, focused on the verdict, has little occasion to draw the following distinction. A moral judgment, in the output sense of Section 1.1, is an evaluative verdict about an action, character, or state of affairs. Moral agency is a status: the standing of a being whose acts can be attributed to it as their source, so that praise, blame, and responsibility apply. In the ordinary human case the two are related, since competent judgment is one expression of agency. But they are not identical, and artificial systems force us to consider that they can come apart. A system might produce impeccable moral verdicts and still fail to be a moral agent, if agency requires something that producing verdicts does not supply.

One might object that there can be no moral judgment without agency, since judging is itself something an agent does. In the act sense of Section 1.1, this is correct, and the paper agrees. In the output sense, it is not: a verdict token can be generated without anyone standing behind it. The paper's thesis is precisely that present AI systems produce moral judgments in the output sense only.

That agency requires more than correct output is a settled theme in moral philosophy, developed independently of any question about machines. To act well is not merely to do what a good person would do. It is to do it for the reasons that make it good, out of responsiveness to those reasons as reasons [27]. The most systematic development of this idea locates moral responsibility in guidance control, which has two components. The mechanism issuing in behavior must be the agent's own, and it must be appropriately responsive to reasons [28]. On such accounts, an agent is a source of conduct not when its outputs happen to be correct, but when they flow from a mechanism that is both its own and reasons-responsive in the right way. A system that reliably outputs the right verdict without being moved by what makes it right satisfies a behavioral criterion but fails the agential one. Reasons-responsiveness of the relevant kind is a condition on being something to which conduct can be imputed at all. Output-matching, however reliable, does not establish it [27,28].

The philosophy of machine ethics has tried to accommodate artificial systems by relaxing this condition rather than meeting it, and this is where the present argument diverges. One influential move distinguishes levels of abstraction at which a system can be described. At a suitable level, it argues, an artificial agent counts as a moral agent—interactive, autonomous, and adaptive—and can be accountable even without the inner states we associate with responsibility [29]. A related taxonomy ranks artificial systems from those with mere ethical impact, through implicit and explicit ethical agents, to the limit case of a full ethical agent [30]. The program of functional morality builds on this. It proposes engineering systems whose behavior is functionally moral, sensitive to morally relevant features and able to act appropriately, without waiting on the metaphysics of consciousness or genuine agency [31].

These proposals are clarifying, and their engineering value is not in dispute. What is in dispute is a semantic drift they encourage. Each secures the term moral agent by redefining agency so that behavioral or functional adequacy suffices for it [29–31]. This is a legitimate stipulation for design purposes, but it concedes the point at issue. If "moral agent" just means "a system whose outputs are functionally moral," a judgment-matching machine qualifies by definition. The question that survives the stipulation is whether functional adequacy captures what we care about when we hold something responsible, trust it, resent it, or forgive it.

The account of the previous sections answers this question on principle rather than by stipulation. What functional accounts leave out is not an unobservable inner glow. It is the self-maintaining stake that turns an output into an evaluation undergone by a being for whom it matters, rather than a classification computed by a mechanism for which nothing does. Output-matching cannot establish this stake, since two systems can produce the same outputs while only one has anything at issue in producing them.

Moral Agency, Moral Standing, and Moral Patienthood

The argument so far concerns moral agency. Agency is not the same as moral standing, and the distinction should be marked explicitly so that the thesis is not taken to prove more than it does. Three notions must be distinguished. Moral judgment is the capacity to produce normative evaluations. Moral agency is the capacity to act from one's own stakes in a way that makes one a source of conduct and a bearer of responsibility. Moral standing, or moral patienthood, is the property of being a locus for which things can go well or badly, so that one's interests make claims on others and one becomes an object of moral consideration. These come apart in familiar cases: an infant or a nonhuman animal may be a moral patient without yet being a full moral agent.

This distinction refines the conclusion rather than weakening it. What the argument establishes directly is a criterion for moral agency: a judgment-matching system that lacks affective-homeostatic self-maintenance is not a moral agent, whatever the quality of its verdicts. Whether such a system has moral standing is a further question. The same framework bears on it without settling it here, since standing and agency rest on overlapping but distinct conditions. If moral standing also depends on there being a being for whom things can go well or badly, then a system lacking the relevant organization would lack standing as well. That conditional is stated here, not defended (for a Spinozist treatment of affect and the standing of artificial systems, see [32]). The paper defends the agency claim and flags the standing question without resolving it.

5. From Sentience to Self-Maintenance

The most forceful philosophical denial that algorithms are moral agents uses a thought experiment from the philosophy of mind. Imagine a being behaviorally indistinguishable from a moral agent but with no first-personal experience at all: a moral zombie. It can do good and evil, but it would neither celebrate a life saved nor regret a life taken, because there is nothing it is like to be it. On this analysis, algorithms are functional moral zombies. They can be neither autonomous nor accountable, and the deep reason is that they lack sentience. Without sentience there is no genuine moral understanding, since grasping what it is to inflict suffering requires experiential acquaintance with suffering [8]. This paper shares the conclusion that judgment-matching machines are not moral agents. The disagreement concerns the ground of that conclusion, and the ground determines whether the argument can make progress.

The worry about resting the case on sentience is not that the criterion is false but that it is misplaced. One might reply that epistemic inaccessibility is no proof of theoretical inadequacy: sentience could be exactly the right criterion even if we cannot detect it from outside. That reply is correct as far as it goes. The present argument does not claim that sentience is irrelevant to moral agency. It makes a more specific claim: sentience is not the explanatory terminus. The thought experiment that motivates the sentience criterion relies on the conceivability of a being whose behavior is complete while its experience is absent [8]. This shows that behavior underdetermines experience. It does not explain why experience should matter to agency in the first place. The account developed here answers that prior question. When phenomenal feeling is morally load-bearing, it is so

because it is the experiential manifestation of a deeper affective-homeostatic organization: the felt aspect of a being's stake in its own persistence [33]. Sentience, so understood, is not rejected. It is repositioned as a symptom of the condition rather than the condition itself.

The intervention this paper proposes is therefore to relocate the ground of agency to affective-homeostatic self-maintenance. What a moral agent has, and a judgment-matching machine lacks, is not merely an inner movie but a stake. A moral agent is a being whose own persistence is at issue in its dealings with the world, whose states register as better or worse relative to that persistence, and whose valuations are undergone rather than merely produced [12,13,19]. This relocation has two advantages over resting the case directly on sentience. First, it is more tractable. Self-maintenance is not a private phenomenal quality but a functional-organizational feature with observable signatures: forms of self-regulation, vulnerability, and endogenous valuation that can be investigated rather than merely asserted or denied. Second, it is better grounded, because it connects moral agency to the origin of value itself rather than to a phenomenal property whose relation to value is left unexplained. Sentience marks where the felt stakes show up; self-maintenance explains why there are stakes to feel.

The distinction that carries the argument is between simulating an output and instantiating a ground. A large language model that reproduces human moral verdicts simulates the output of a moral agent. It does not instantiate the self-maintaining ground from which, in the human case, that output derives its point. This is not a claim about substrate. The argument does not say that only carbon can value, or that self-maintenance is impossible in principle for an engineered system. A system genuinely organized around its own persistence, with real endogenous stakes, would satisfy the criterion whatever it were made of. The claim is narrower and therefore defensible: present-day judgment-matching systems are not so organized. What they have is a statistical command of the linguistic and behavioral traces of beings who do have such a stake. They have learned the shadow that valuation casts in language, and they cast it back with uncanny accuracy. That is why the convergence is real. But a shadow, however faithful, is not lit from within.

5.1. Endogenous Self-Maintenance Versus Externally Regulated Systems

If self-maintenance is the criterion, a hard question follows. A thermostat maintains temperature, a control system regulates its variables, and countless engineered devices hold themselves within set bounds. Are these moral agents, or even candidates? They are not, and explaining why sharpens the criterion into something usable. Not every self-regulating system instantiates the relevant kind of self-maintenance; in enactive accounts, genuine agency requires that the norms of a system's activity be generated by its own self-constituting organization [24,34]. In the present framework, homeostatic self-maintenance grounds agency only when it meets four conditions:

1. It is endogenous: the norms it maintains arise from the system's own organization rather than being imposed by a designer.
2. It is organizationally integrated: self-maintenance is constitutive of the system's persistence, not one function among others.
3. It is affectively valenced: departures from viability register as better or worse for the system.
4. It is action-guiding: these registered stakes can move the system's conduct.

A thermostat meets none of these conditions in the required sense. Its set point is stipulated from outside. Its regulation is not constitutive of its being a thermostat. Nothing registers as mattering to it. And its stakes are its designer's, not its own.

This is also the place to be precise about a claim that could otherwise overreach. It is tempting to say that nothing helps or harms a present-day AI system. But, as the

functionalist rightly notes, such systems can be degraded by loss of power, hardware failure, memory corruption, or shutdown. The point is not that no state of the system can change for the worse. It is that such changes are defined from the standpoint of designers and users, not from an endogenous normative stake of the system's own. A shutdown harms the interests that deployed the system, not interests the system maintains for itself, because it maintains none. External regulation is not endogenous self-maintenance. The decisive contrast, then, is not the presence or absence of regulation; AI systems are regulated in elaborate ways. It is the source of the regulation: whether the norms in play arise from the system's own self-maintaining organization or are imposed from without.

5.2. What Current Architectures Do and Do Not Have

The claim that present systems lack endogenous valuation must be stated carefully, because some current architectures contain components that look like valuation. Three cases deserve attention.

First, large language models are not trained on text prediction alone. After pretraining, they are typically fine-tuned with reinforcement learning from human feedback. A reward model, trained on human rankings of model outputs, scores candidate outputs, and the language model is optimized to produce high-scoring ones [35]. In a functional sense, such a system has a preference ordering over its outputs. But this ordering is scoring, not valuation in the sense of Section 1.1. Its norms are induced from human labelers' judgments and fixed by developers. The optimization takes place during training, not as the system's ongoing regulation of its own persistence. And in typical deployment the parameters are fixed, so nothing the model outputs makes its own condition better or worse. The reward model therefore reinforces the point of Section 6: it is one more channel through which human valuations are compressed into the system.

Second, reinforcement-learning agents act in environments so as to maximize a reward signal. Here the stake seems closer, since the agent's own actions change the reward it receives. Yet in standard designs the reward function is specified exogenously, and the agent's continued existence does not depend on it. An agent that performs badly is retrained or replaced; it does not decline toward its own dissolution. Its reward is the designer's measure of success, imposed from outside. In the terms of Section 5.1, this is external regulation.

Third, and most interesting, are research programs that build homeostasis into artificial agents. Homeostatic reinforcement learning defines reward in terms of the reduction in deviations from internal set points and shows that reward-seeking and physiological stability then coincide [36]. Proposals in soft robotics go further. They envision machines with vulnerable bodies whose internal states must be regulated for the machine to persist, and they suggest that such homeostatic design may yield equivalents of feeling [37]. The present framework explains why these programs are the right place to look for future candidates: they target exactly the condition that output-matching ignores. Even so, current implementations meet the four conditions of Section 5.1 only partially. Their set points are still chosen by designers, and the regulated variables are typically simulated rather than constitutive of the system's material persistence.

The paper's claim is accordingly restricted. Present-day language models, including those fine-tuned with human feedback, and standard reinforcement-learning agents lack endogenous valuation. They fall short not marginally but categorically, and at the first, generative level identified in Biological Normativity Is Not Yet Moral Normativity Section. They do not merely lack moral normativity; they lack the endogenous valuation from which any normativity for the system could begin. Homeostatically designed agents approach

the criterion, but they have not yet met it. Whether they will is an open empirical and engineering question, which the criterion makes precise rather than closes.

6. The Convergence Illusion

We can now reinterpret the empirical finding with which the paper began. The reinterpretation is deflationary without being dismissive. Machines increasingly reproduce human moral verdicts [1–3], and people's judgments of AI and human agents can be studied within a common experimental frame [4]. These are genuine and important results. Given the argument of the previous sections, however, they do not and cannot show that machines are approaching moral agency. The convergence illusion is the inferential error of treating convergence in moral judgment as evidence of convergence in moral agency. It is an illusion because the two vary independently. A system can approach perfect convergence on the output while remaining at zero on the ground, since the informational content of valuation can be modeled independently of the organization that makes anything matter to the valuer.

The illusion has a specific mechanism. Systems that reproduce human moral judgments do so by capturing regularities in a vast record of human moral discourse [2]. That record was produced, in every case, by beings who did have a homeostatic stake in their affairs. The value-laden structure of this discourse is the compressed residue of countless episodes of real valuation by real self-maintaining creatures. As Section 5.2 noted, preference-based fine-tuning adds a further layer of the same residue [35]. A model that predicts this structure with high fidelity is not thereby valuing. It is reconstructing the statistical form of others' valuing. This is why the outputs can be so good while the ground is absent.

The same mechanism explains a feature of the results that would otherwise be puzzling. In ambiguous moral scenarios, some models' choices are sensitive to morally irrelevant changes in question wording [38], and the moral judgments encoded in language models reflect the biases of the human text on which they were trained [2]. A system that reconstructs the form of valuation from traces will follow those traces wherever they lead, including into their distortions. It has no independent stake against which to check them. A being that values from a homeostatic ground has such a check, however imperfect. Some outcomes are actually better or worse for it, and this supplies a friction that pure trace-reconstruction lacks.

The reinterpretation cuts in two directions, and both matter. Against the over-reading of convergence, it warns that judgment-matching is not agency-approximation, and that no accumulation of the former yields the latter. But it also vindicates and extends the empirical program. People judge an AI's morally identical act differently from a human's [4,15,16]. From the present vantage point, this looks less like a bias to be corrected than like an implicit recognition of the asymmetry defended here. Folk moral cognition may be tracking, however inarticulately, the difference between an act that issues from a being with a stake and one that issues from a system without one. This is a hypothesis that the empirical program is well placed to test, and Section 8 states it as such.

7. Objections and Replies

The argument invites several strong objections. Meeting them squarely is the best test of its durability.

7.1. "Affect Is Already in the Model as System 1"

The first objection comes from within the dual-process tradition. Affect, it says, is already in the model as System 1; the paper has relabeled a familiar component and dressed it in metaphysics. This misreads the claim. As Section 2 stressed, the routing assumption is

a feature of how affect is operationalized, not a doctrine every dual-process theorist holds. The thesis here is that affect, in its foundational role, is not a verdict-producing process but the ground of stake from which verdicts derive their status as moral [5,19]. In the terms of Section 1.1, System 1 concerns affect as a contributor to judgment in the output sense, while the present thesis concerns affect as the registering of valuation. These are claims at different levels, not rival labels. The framework's own difficulties suggest that the distinction is not idle. Suppose the affective/cognitive partition were as clean and complete an inventory of morality's ingredients as the operationalization implies. Then the recurrent trouble in replicating some of its reported process-level effects [17], and the charge that its core assumptions are insulated from test [18], would be harder to understand.

7.2. The Is–Ought Objection

The second objection targets the Spinozan foundation and, through it, the paper's argument. Spinoza is standardly charged with deriving an "ought" from an "is": from the fact that each thing strives to persist, he infers that the good is what aids that striving. The worry echoes Hume's classic complaint against moving from claims about what is to claims about what ought to be [39] (T 3.1.1.27). The inference also seems open to counterexample. Consider a hermit who meditates on a mountain and eats very little. His conduct does not serve his self-preservation, yet it seems clearly good. If goodness can come apart from self-preservation, the objection runs, self-preservation cannot ground moral agency. The point also seems to transfer to AI: a system might likewise do something good that serves no self-maintaining interest. There are three replies.

First, the argument of this paper does not move from "is" to "ought." It makes a constitutive claim about agency, not a normative claim about goodness. It states a necessary condition for a capacity—what a being must be like in order to be a moral agent—and is silent on what makes acts right or good. "Moral agency presupposes a self-maintaining stake" is a descriptive claim about the conditions of a status. In the same way, "promising presupposes language" is a descriptive claim about the conditions of a practice. No ought is derived, and Hume's gap stays where it was. The paper therefore neither needs nor adopts the stronger Spinozan identification of the good with what preserves the agent. It borrows from Spinoza only the structural thesis that valuation originates in the striving of a being whose persistence is at issue. Biological Normativity Is Not Yet Moral Normativity Section already marks the gap between this biological normativity and moral normativity and leaves open how the latter arises.

Second, the hermit is less of a counterexample to Spinoza than he appears. Conatus is not a drive to maximize biological survival. It is a striving to persevere in one's being, and what the mind strives for insofar as it reasons is understanding. Spinoza accordingly holds that the mind's greatest good is knowledge of God, and that blessedness is not the reward of virtue but virtue itself [19] (E4P26–P28, E5P42). A contemplative life that gives up bodily comfort for understanding is, in Spinoza's terms, an increase in the mind's power, not a failure of self-preservation. On this reading the ascetic is a candidate exemplar of conatus rather than a refutation of it. Extreme cases, in which contemplation is pursued to the point of destroying the body, remain contested in Spinoza scholarship. The present argument does not need to settle them, because it does not rest on Spinoza's theory of the good.

Third, and most important for the thesis, the hermit shows how moral significance depends on stake. His abstinence has moral weight because he has something to lose. The hunger, cold, and frailty are his own, and his renunciation is the voluntary subordination of a stake he genuinely has. Self-denial, sacrifice, and courage all presuppose a self whose interests are denied, sacrificed, or risked. Applied to AI, this cuts against the objection. An AI system that "abstains" from consuming computational resources renounces nothing,

because nothing is at issue for it. Its abstention may have good effects, and the paper fully grants that AI outputs can be good. What the system cannot do is give anything up. The hermit therefore shows that an act's goodness is not reducible to its contribution to the agent's survival, a point the paper accepts. At the same time, he confirms that moral agency requires a being whose own persistence is at issue. Far from being a counterexample, he is one of the clearest cases of the asymmetry the paper defends.

7.3. The Functionalist Objection

The third objection is the functionalist's, and it is the most serious. Suppose a future system were organized to have genuine endogenous stakes: self-maintaining, vulnerable, valuing on its own account. By the paper's own criterion, it would be a moral agent. So, the criterion concedes that machine agency is possible and merely bets, contingently, that current systems do not meet it. This is correct. The argument is not a proof that artificial moral agency is impossible, and it is not carbon chauvinism. The criterion is substrate-neutral, turning on self-maintaining organization rather than on the material that realizes it. Section 5.1 states the four conditions any candidate must meet, and Section 5.2 identifies the research programs that come closest [36,37]. The claim about present systems is empirical and defeasible, and the paper welcomes this. A criterion that could be met is a criterion that says something. A criterion resting on an undetectable property can be neither met nor failed in any way we could recognize [8]. Locating agency in self-maintenance tells us what to look for. It does not decree what we will find.

7.4. The Responsibility-Vacuum Objection

The fourth objection comes from moral practice. Autonomous systems already make decisions with life-and-death consequences. The standard illustration is the crash-optimization dilemma of self-driving vehicles, on which public intuitions have been mapped at scale [40,41]. Work on "responsibility gaps" created by learning machines [42] and autonomous weapons [43] sharpens the worry. If such systems are not moral agents, does the argument leave a responsibility vacuum exactly where responsibility matters most?

On the contrary, it locates the vacuum correctly. The temptation to treat the machine as a locus of responsibility is itself a symptom of the convergence illusion: because the system produces morally weighty outputs, we are drawn to treat it as answerable for them. Denying it agency does not leave the outputs unowned. It returns responsibility to the designers, deployers, and institutions whose choices the system executes. The account does not close every gap by itself. When designers could not foresee a learning system's behavior, assigning responsibility to them raises real questions [42]. But the account shows that this is a gap among human agents and institutions, to be closed by institutional design, not by promoting the machine to agency. Recognizing that a vehicle is a moral judge without moral agency excuses no one. It prevents the machine from serving as a scapegoat that absorbs a responsibility no mechanism can bear.

7.5. "Why Spinoza and Damasio?"

The fifth objection questions the philosophical machinery. Why route the argument through Spinoza and Damasio, when the distinction between competent output and genuine agency could be stated in neutral terms? There are two reasons. First, a neutral statement without an account of why the ground matters is easily dismissed as intuition-mongering. Second, the sentience-based alternative that does try to supply a ground founders on inaccessibility [8]. The conatus–homeostasis framework derives the criterion from an independently motivated account of where value comes from. It converges, from seventeenth-century metaphysics and twenty-first-century neurobiology, on the same struc-

tural claim [12,19], while keeping the two levels of description distinct, as Section 3 insisted. The apparatus is not ornament. It turns a bare intuition that machines lack something into a specified thesis about what they lack and why it matters.

7.6. "So What?"

A final objection grants everything and asks what follows. If machines are moral judges without moral agency, so what, as long as their judgments are good? The distinction matters wherever we do more with a moral verdict than consult it. We hold agents responsible, extend and withdraw trust, resent and forgive, and treat some beings as having claims on us. Each of these practices presupposes an agent with a stake, not merely a source of verdicts [27,28]. Knowing that a system judges without being an agent tells us that it may be consulted but not blamed, and used but not trusted as we trust a person. It also tells us not to treat it as bearing the moral standing whose outputs it so convincingly reproduces. This is not a small consequence. It is the difference between a tool that speaks the language of morality and a member of the moral community.

8. Implications: A Behavioral Research Program

The thesis is conceptual, but it has direct consequences for empirical work: it specifies a research program. The debate over machine moral agency has oscillated between two kinds of criteria. Behavioral or functional criteria make agency too cheap [29–31]. Sentience-based criteria make it undetectable [8]. Affective-homeostatic self-maintenance lies between them. It is demanding enough to withhold agency from mere judgment-matching yet specified enough to be investigated. Adopting it reframes a familiar question. Rather than asking whether machines can judge as we do—increasingly, they can—we should ask what people's attributions of moral agency track. Do they track judgment accuracy, or cues of endogenous stake? The two can be pulled apart experimentally.

It should be stressed what such a program can and cannot show. It tests folk attributions, not the metaphysical thesis itself. But it can show whether ordinary practices of responsibility and trust track the property the thesis identifies, and that question matters both for the theory and for the design and governance of AI systems.

8.1. Hypotheses

The account yields five testable predictions.

- **H1.** *Increasing a system's moral-judgment accuracy will increase perceived moral competence without necessarily increasing perceived moral agency.*
- **H2.** *Cues of self-maintenance, that is, signs that a system has endogenous stakes, will increase moral-agency attribution independently of judgment accuracy.*
- **H3.** *Cues of vulnerability and self-concern will increase moral-standing attribution more strongly than competence cues do.*
- **H4.** *High judgment accuracy without self-maintenance cues will produce a recognizably competent but non-agentic attribution profile: people will credit the system's verdicts while withholding responsibility from it.*
- **H5.** *Describing a system in terms of endogenous stakes will increase agency attribution more than describing it in terms of computational sophistication.*

8.2. Operationalization: Three Illustrative Studies

The following designs show how these hypotheses can be tested with standard methods. They are offered as templates, not final protocols.

Study 1 (H1, H2, H4): Dissociating accuracy from stake. A preregistered 2 × 2 between-subjects vignette experiment crosses judgment accuracy with self-maintenance

cues. In the accuracy manipulation, participants review an AI system's verdicts on ten standard moral scenarios. In the high-accuracy condition, the verdicts match the normative consensus established in a pretest; in the low-accuracy condition, they match it on six of ten items. In the self-maintenance manipulation, the system is described either as regulating its own energy supply and physical integrity, with behavior that changes when these are threatened, or as a system whose operation is regulated entirely by its operators. The two descriptions are matched for length and technical detail. Dependent measures include perceived moral competence (e.g., "This system evaluates moral situations correctly") and perceived moral agency, measured with a validated instrument such as the Perceived Moral Agency Scale [44]. A responsibility-allocation task is also included: participants read a scenario in which the system's decision causes harm and divide 100 points of responsibility among the system, its developers, and the deploying organization. H1 predicts a main effect of accuracy on competence but not on agency. H2 predicts a main effect of self-maintenance cues on agency and on the share of responsibility assigned to the system. H4 predicts that the high-accuracy/no-cue cell will show high competence together with low agency and low responsibility allocation, which can be tested with planned contrasts.

Study 2 (H3, H5): Framing stake versus sophistication. A single system is described under three framings. An endogenous-stake framing says the system monitors its own reserves and protects itself from damage that would end its operation. A computational-sophistication framing emphasizes model size and reasoning ability. A neutral control describes the system's function only. A separate factor adds or omits a vulnerability cue (e.g., the system can be irreversibly damaged). Agency and experience are measured with the two dimensions of mind perception [45]. Moral standing is measured by the judged wrongness of damaging or permanently shutting down the system and by participants' willingness to protect it. H5 predicts higher agency attribution under the endogenous-stake framing than under the sophistication framing. H3 predicts that the vulnerability cue will raise standing attributions more than the sophistication framing does.

Study 3 (Section 6): Extending the human–AI asymmetry. This study extends the trolley and footbridge paradigm of Zhang et al. [4] by adding a third agent type: an AI described with the self-maintenance cues of Study 1. If folk moral cognition tracks stake, as Section 6 proposes, wrongness and blame ratings for the self-maintaining AI should move toward those for the human agent, relative to the standard AI. A behavioral extension can use an incentivized advice-taking paradigm. Participants receive advice on moral dilemmas from human and AI advisors. In one condition, the AI advisor's verdicts change under morally irrelevant rewordings, mirroring the wording sensitivity documented by Scherrer et al. [38]. Advice weighting then provides a behavioral measure of trust that complements self-report.

Across studies, manipulation checks should confirm that the cues are perceived as intended, and a priori power analyses should determine sample sizes. The studies manipulate descriptions of systems, not their actual organization, which is a limitation. But this is exactly what folk attribution responds to, and it is the variable that public communication about AI systems also manipulates.

8.3. Practical Implications

Two practical implications follow from the same argument. First, attributions of moral agency to machines are category errors with real costs. The chief cost is the diffusion of responsibility away from human actors toward systems that cannot bear it. Recognizing machines as instruments of human moral agency keeps responsibility with the humans in the loop. Second, in domains where machine verdicts carry weight, the operative question is not whether the machine is a good moral agent. It is whose stake its output serves and

who answers for it. The conceptual claim and the empirical program are not competitors. The former specifies what the latter should look for once it stops treating convergence as the end of the story.

9. Conclusions

Machines have learned to judge as we judge, and it is tempting to conclude that they are becoming, in some thin sense, moral. This conclusion mistakes an output for a ground. Moral judgment, in the sense that empirical work measures, is a verdict. Moral agency is the standing of a being for whom the matter judged is genuinely at stake. That standing rests on affective-homeostatic self-maintenance: the conatus-like striving from which valuation springs, which is Spinoza's actual essence of a thing and, in a distinct register, Damasio's felt homeostasis. This ground is offered as necessary and constitutive, not as a sufficient theory of the moral. As Biological Normativity Is Not Yet Moral Normativity Section argued, biological normativity must still become the moral kind, and the paper marks that threshold without fully crossing it. The claim is also constitutive rather than normative. It specifies what a moral agent must be, not what is good, and so it does not infer an ought from an is (Section 7.2). But because the ground can be absent while its outputs are reproduced, human–AI convergence in moral judgment is fully compatible with the absence of moral agency in the machine. The convergence is not an illusion; the inference from convergence to agency is.

The argument's most consequential move is to relocate the criterion of agency from sentience to self-maintenance. It does not reject sentience. It repositions it as the experiential manifestation of a deeper organization, so that what looked like the explanatory terminus becomes a symptom of the condition. A sentience-first criterion leaves the debate stalled on a property no observation can reach. The self-maintenance criterion, by contrast, is functional, investigable, and continuous with the best account we have of where value comes from. It can be tested against current architectures (Section 5.2), and it turns an unfalsifiable stalemate into the research program of Section 8.

In the meantime, the argument asks for a certain restraint. We should receive the machine's fluent moral speech as what it is: the accurate shadow of valuations performed by beings with something to lose. And we should keep responsibility, trust, and moral standing where a stake actually resides. Intelligence was never enough for consciousness, and judgment, however accurate, is not enough for moral agency. What is missing in both cases is the same: a being for whom things can go well or badly, and for whom, therefore, the difference between right and wrong is not a computation but a concern.

Funding: This research was funded by the Scientific Research Projects Coordination Unit (BAP) of İnönü University, grant number SBA-2025-4264.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The author declares no conflicts of interest.

Note

[1] References to the Ethics follow the standard convention: E = Ethics, followed by part number; D = definition; P = proposition; Dem = demonstration; S = scholium.

References

1. Dillion, D.; Tandon, N.; Gu, Y.; Gray, K. Can AI language models replace human participants? *Trends Cogn. Sci.* **2023**, *27*, 597–600. [CrossRef] [PubMed]
2. Schramowski, P.; Turan, C.; Andersen, N.; Rothkopf, C.A.; Kersting, K. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nat. Mach. Intell.* **2022**, *4*, 258–268. [CrossRef]
3. Aharoni, E.; Fernandes, S.; Brady, D.J.; Alexander, C.; Criner, M.; Queen, K.; Rando, J.; Nahmias, E.; Crespo, V. Attributions toward artificial agents in a modified Moral Turing Test. *Sci. Rep.* **2024**, *14*, 8458. [CrossRef] [PubMed]
4. Zhang, Y.; Wu, J.; Yu, F.; Xu, L. Moral judgments of human vs. AI agents in moral dilemmas. *Behav. Sci.* **2023**, *13*, 181. [CrossRef] [PubMed]
5. Greene, J.D.; Sommerville, R.B.; Nystrom, L.E.; Darley, J.M.; Cohen, J.D. An fMRI investigation of emotional engagement in moral judgment. *Science* **2001**, *293*, 2105–2108. [CrossRef] [PubMed]
6. Haidt, J. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychol. Rev.* **2001**, *108*, 814–834. [CrossRef] [PubMed]
7. Greene, J.D. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*; Penguin Press: New York, NY, USA, 2013.
8. Véliz, C. Moral zombies: Why algorithms are not moral agents. *AI Soc.* **2021**, *36*, 487–497. [CrossRef] [PubMed]
9. Coeckelbergh, M. Robot rights? Towards a social-relational justification of moral consideration. *Ethics Inf. Technol.* **2010**, *12*, 209–221. [CrossRef]
10. Gunkel, D.J. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*; MIT Press: Cambridge, MA, USA, 2012.
11. Damasio, A.R. *Descartes' Error: Emotion, Reason, and the Human Brain*; Putnam: New York, NY, USA, 1994.
12. Damasio, A. *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain*; Harcourt: Orlando, FL, USA, 2003.
13. Damasio, A.; Carvalho, G.B. The nature of feelings: Evolutionary and neurobiological origins. *Nat. Rev. Neurosci.* **2013**, *14*, 143–152. [CrossRef] [PubMed]
14. Colombetti, G. *The Feeling Body: Affective Science Meets the Enactive Mind*; MIT Press: Cambridge, MA, USA, 2014.
15. Bigman, Y.E.; Gray, K. People are averse to machines making moral decisions. *Cognition* **2018**, *181*, 21–34. [CrossRef] [PubMed]
16. Malle, B.F.; Scheutz, M.; Arnold, T.; Voiklis, J.; Cusimano, C. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15), Portland, OR, USA, 2–5 March 2015*; ACM: New York, NY, USA, 2015; pp. 117–124.
17. Zou, L.; Wu, J.; Liu, C. No detectable effects of chronic stress on moral decision-making are found: A data reanalysis and a pre-registered replication and extension. *Behav. Sci.* **2025**, *15*, 1068. [CrossRef] [PubMed]
18. Quinlan, P.T.; Cohen, D.J. Untested assumptions and tenuous evidence: A critique of the dual-process account of moral judgment. *Behav. Sci.* **2026**, *16*, 20. [CrossRef] [PubMed]
19. Spinoza, B. Ethics. In *The Collected Works of Spinoza*; Curley, E., Ed. and Translator; Princeton University Press: Princeton, NJ, USA, 1985; Volume 1.
20. Deleuze, G. *Spinoza: Practical Philosophy*; Hurley, R., Translator; City Lights Books: San Francisco, CA, USA, 1988.
21. Damasio, A. *The Strange Order of Things: Life, Feeling, and the Making of Cultures*; Pantheon: New York, NY, USA, 2018.
22. Jonas, H. *The Phenomenon of Life: Toward a Philosophical Biology*; Harper & Row: New York, NY, USA, 1966.
23. Thompson, E. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*; Belknap Press of Harvard University Press: Cambridge, MA, USA, 2007.
24. Di Paolo, E.A. Autopoiesis, adaptivity, teleology, agency. *Phenomenol. Cogn. Sci.* **2005**, *4*, 429–452. [CrossRef]
25. Nussbaum, M.C. *Upheavals of Thought: The Intelligence of Emotions*; Cambridge University Press: Cambridge, UK, 2001.
26. Prinz, J.J. *The Emotional Construction of Morals*; Oxford University Press: Oxford, UK, 2007.
27. Purves, D.; Jenkins, R.; Strawser, B.J. Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory Moral Pract.* **2015**, *18*, 851–872. [CrossRef]
28. Fischer, J.M.; Ravizza, M. *Responsibility and Control: A Theory of Moral Responsibility*; Cambridge University Press: Cambridge, UK, 1998.
29. Floridi, L.; Sanders, J.W. On the morality of artificial agents. *Minds Mach.* **2004**, *14*, 349–379. [CrossRef]
30. Moor, J.H. The nature, importance, and difficulty of machine ethics. *IEEE Intell. Syst.* **2006**, *21*, 18–21. [CrossRef]
31. Wallach, W.; Allen, C. *Moral Machines: Teaching Robots Right from Wrong*; Oxford University Press: Oxford, UK, 2009.
32. Gültekin, A. Can an algorithm be sacred? Affect, embodiment, and the spiritual status of artificial intelligence: A Spinozist, cross-traditional inquiry. *Religions* **2026**, *17*, 1049. [CrossRef]
33. Gültekin, A. Why intelligence is not enough: Homeostasis, affect, and the biological ground of consciousness. *Sci. Cult.* **2026**, *12*, 6377–6382.
34. Barandiaran, X.E.; Di Paolo, E.; Rohde, M. Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adapt. Behav.* **2009**, *17*, 367–386. [CrossRef]

35. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744. [CrossRef]
36. Keramati, M.; Gutkin, B. Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife* **2014**, *3*, e04811. [CrossRef] [PubMed]
37. Man, K.; Damasio, A. Homeostasis and soft robotics in the design of feeling machines. *Nat. Mach. Intell.* **2019**, *1*, 446–452. [CrossRef]
38. Scherrer, N.; Shi, C.; Feder, A.; Blei, D.M. Evaluating the moral beliefs encoded in LLMs. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 51778–51809. [CrossRef]
39. Hume, D. *A Treatise of Human Nature*; Norton, D.F., Norton, M.J., Eds.; Oxford University Press: Oxford, UK, 2000.
40. Bonnefon, J.-F.; Shariff, A.; Rahwan, I. The social dilemma of autonomous vehicles. *Science* **2016**, *352*, 1573–1576. [CrossRef] [PubMed]
41. Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; Rahwan, I. The Moral Machine experiment. *Nature* **2018**, *563*, 59–64. [CrossRef] [PubMed]
42. Matthias, A. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics Inf. Technol.* **2004**, *6*, 175–183. [CrossRef]
43. Sparrow, R. Killer robots. *J. Appl. Philos.* **2007**, *24*, 62–77. [CrossRef]
44. Banks, J. A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Comput. Hum. Behav.* **2019**, *90*, 363–371. [CrossRef]
45. Gray, H.M.; Gray, K.; Wegner, D.M. Dimensions of mind perception. *Science* **2007**, *315*, 619. [CrossRef] [PubMed]

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.