Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Toward a Kantian Account of AMA through Practical Apperception

Sung-Yeop Jo

sungyeop.jo@sejong.ac.kr

Sejong University

Research Article

Keywords:

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11399646/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Toward a Kantian Account of AMA through Practical Apperception

I. Introduction

Can current AI systems, which now seem to match—or even surpass—average human performance on moral-reasoning tasks, be regarded as moral agents? Since Floridi and Sanders' seminal paper (Floridi & Sanders, 2004), the debate over the concept and possibility of artificial moral agents (AMAs) has attracted sustained scholarly attention. Related works in machine and robot ethics are now widely regarded as pioneering contributions that helped lay the groundwork for the contemporary reconsideration of this concept,[1] which has intensified since the public release of ChatGPT in 2022.

The aim of this paper is to contribute to the contemporary AMA debate using conceptual resources of Kantian philosophy. I contend that current LLM-based conversational systems (LCSs), such as ChatGPT and Claude, cannot be regarded as genuine moral agents under the Kantian conditions for agency insofar as they are *provisionally* judged not to possess the capacity to autonomously integrate competing incentives into a unified principle of action—a capacity that I designate as *practical apperception*.[2] An epistemic difficulty in assessing artificial moral agency lies in the fact that *I* cannot directly access the minds of others, let alone the minds of AI systems—if such minds exist—from *my* first-personal perspective. Kant's philosophy of mind accommodates this first-personal limitation and offers a detour by permitting the provisional recognition of others as minded subjects.

In what follows, I first reconstruct Kant's theory of apperception in Section II and then apply it in Section III to defend the contention that LCSs, at least in their current form, do not satisfy a necessary Kantian condition for moral agency. I thereby support a skeptical assessment of their moral potential without advancing ontological claims about their inner states.

[1] Among many others, see: (Coeckelbergh, 2009, 2010; Gunkel, 2018; Sparrow, 2007)

[2] By LCSs, I understand current, general-purpose, publicly deployed LLM-based conversational systems, exemplified by ChatGPT and Claude in their ordinary conversational modes. In this article, I focus on this particular class of AI systems because it constitutes a especially revealing critical case. Current LCSs produce first-personal avowals, reasons for action, apparent deliberation, commitments, revisions, and responses to normative challenges—the very performances most likely to invite the attribution of a practical standpoint.

II. Revisiting Kant's Philosophy of Mind in the Era of Artificial Intelligence

1 A Terminological Concern: Agent and Subject

Before turning to textual analysis, it warrants briefly noting that Kant himself does not use the terms "agent" or "agency," nor do these terms have straightforward German equivalents that fully capture their philosophical implications, although his philosophy of mind has commonly been recognized as relevant to the contemporary understanding of these concepts. What stands at the center of Kant's analysis, rather, is the *subject*. The attempt to bring the resources of Kantian philosophy to bear on the contemporary debate over AI agency can nonetheless be justified insofar as, for Kant, subjectivity constitutes a necessary condition for agential capacities. Kant's analysis of apperception, which I address in this section, characterizes human beings as subjects by identifying the necessary ground to which the mind's cognitive and practical acts can be ascribed. The inference to a subject underlying such acts—say, a good deed—is justified insofar as such acts must be grounded in *agential* capacities, whose bearer counts as an independent *subject*. That being said, although Kant does not explicitly identify a human subject as an agent, his argument justifies positing a subject, to whom agential capacities are ascribed, on the explanatory ground that these capacities cannot stand without an underlying subject. The Kantian subject thereby occupies a necessary placeholder role in the analysis of such acts. In this regard, Kant's Deductions adopt a strategy analogous to that of the Cartesian *cogito* in that they depart from mental activities that cannot be explained without being ascribed to a subject.

Since this paper aims to develop a *necessary* Kantian condition for artificial moral agency, whose nonfulfillment entails the denial of AMA status, adopting a broader sense of Kant's "subject" helps avoid unnecessary complications: If agency requires subjectivity, then showing that LCSs cannot be Kantian *subjects* is sufficient to establish that they cannot be *agents* in the Kantian sense either. Kantian skepticism about AMAs, discussed below, also rests on the claim that the AI models under consideration cannot be recognized as subjects. I therefore restrict the scope of this paper to defending the following thesis: Unless LCSs can be taken to possess a capacity of their own that can be characterized as practical apperception, they are not to be recognized as Kantian subjects and, accordingly, as AMAs.

2 Kant on Apperception(s)

Recognizing current LCSs as AMAs entails the claim that their operations can be regarded as expressions of agential capacities ascribed to the systems themselves as bearers of these capacities. I do not regard my laptop, for example, as a co-author of my writings, although they could not have taken their present form without its tireless assistance. The fact that current LCSs are as entities whose contributions can even be *plagiarized* indicates their distinctive position in our conceptual universe of agents. The recognition of LCSs as a class of agents, however, cannot simply be taken for granted: AMAs would bear, if there were indeed such things, responsibilities for their actions insofar as those actions would be ascribed to them, unlike my laptop, which is not held accountable for the writings to which it contributed. Human moral agency presupposes the *self* to which actions and responsibility for them are ascribed. Artificial moral agency is therefore precluded unless candidate systems can be regarded as self-like subjects.[3]

Do current LCSs have a self? Kant's theory of apperception provides valuable conceptual resources for assessing the contemporary debate over artificial selfhood. Kant's Transcendental Deduction (TD) addresses the ground of my first-personal self-recognition as an *I*. Despite the abstruseness of his terminology, basic ideas of the TD are not difficult to grasp: How can it be justified to posit an *I* as the subject underlying *my* diverse activities? Kant's apperception argument in the TD takes as its starting point the synthesis of sense perceptions, which are distinct in themselves, into a unified representation of an object: Diverse perceptions acquired through sense organs, such as color (red), shape (round and spherical), tactile qualities (firm and cold), and smell (a sweet aroma), would remain separate without forming a single representation of an apple. The *concept* of an apple provides a formal basis for the combination of these perceptions. In Kant's own words, the concept serves as a *rule* for combining sense perceptions: A person who understands the concept of an apple can recognize a sweet smell as belonging to the same red, spherical object in front of her because apples have, as she knows, such a smell.[4] The synthesis of these perceptions into a unified representation thus requires the active participation of a cognitive subject—namely,

[3] I thereby reject the less demanding standard of moral agency that Moor proposes (Moor, 2006). Entities falling into his low-level categories of agency, such as "ethical impact agents" and "implicit ethical agents," cannot be held responsible for their own choices and therefore not genuine agents at all; at best, they are mere aids to agential decision-making—just as my laptop helps me write.

[4] A less intuitive illustration, which Kant himself introduces in the A-Deduction, is the representation of a *body* (CPR A106). The concept of a body, as Kant understands it, has diverse marks, such as extension, impenetrability, shape, and so forth. These marks serve as rules for judging something to be a body. To put the point in terms of the apple example, one does not say that X is a body, if X has no extension. On the basis of the conceptual marks of a body, a cognitive subject can thus identify something as bodily by distinguishing it from non-bodily entities in her surroundings—just as one does not attribute to an apple properties that do not belong to its concept, such as loudness, even when one eats it in a loud room.

a self to whom the objective cognition is ascribed as *her* own; without such cognitive contribution, sense perceptions would remain a mere aggregate rather than a unified manifold.[5] Kant's formulation of the "I think," which "must be able to accompany all my representations" (CPR B132),[6] articulates the cognitive ground of human selfhood: Empirical representations of objects, formed through the conceptual combination of distinct sense perceptions, can count as *mine* because they are unified by the thinking "I," which gives them a cognizable form through synthesis under concepts as rules.

The cognitive ascription of an objective representation to the thinking *I* gives rise, from a practical perspective, to the corresponding question of how an action can be imputed—that is, practically attributed—to an agent. Building on the work of Kant scholars who identify a systematic need to locate the bearer of moral agency corresponding to transcendental apperception,[7] I address the bearer of practical selfhood under the heading of *practical apperception*—a term that refers to the mental capacity to unite *diverse* and even competing incentives into a *single* determination of the will. Human beings experience desires, which become incentives when they function as motives for action. I can feel a craving for sweets, for example, despite my doctor's urgent advice to refrain from eating them. In this case, my decision would result from a free choice either to follow or to resist the craving. Although these two incentives—namely, one directed toward gustatory pleasure and the other toward health—stem from different sources, they can compete and be ordered through an agential act to take them up within a *single* practical standpoint, as expressed in judgments such as: "I desire health even more than this momentary pleasure" or "I value this pleasure more than my health."[8] A more Kantian illustration of the wavering will would involve a decision with moral implications: Kant introduces the case of an honest man, who is threatened with execution by a corrupt ruler attempting to secure his false testimony against another innocent person (CPrR 5:30). Despite the immense pressure, this man—as we too would, were we in his position—can recognize both that he ought not succumb to the threat and that he is, at least in principle, capable of resisting it. That being said, this man can *choose* either to give in to the demand of the corrupt ruler or to remain truthful. His

[5] I do not engage here with the detailed interpretive issues concerning Kant's Deductions, since addressing them falls outside the scope of this paper. In my view, following works provide the most tenable reconstructions of Kant's A-Deduction: (Grüne, 2009, pp. 149–192; Longuenesse, 1998, pp. 35–58)

[6] References to Kant's works use the following abbreviations: *CPR*, *Critique of Pure Reason* (Kant, 1998); *CPrR*, *Critique of Practical Reason* (Kant, 2015); *CJ*, *Critique of the Power of Judgment* (Kant, 2000); and *Religion*, *Religion within the Boundaries of Mere Reason* (Kant, 1996). *CPR* is cited according to the standard A/B pagination.

[7] For attempts to reconstruct the concept of practical apperception that Kant himself leaves largely unexplained, see: (Hoffmann, 2002, pp. 426–433; Jo, 2026, pp. 33–66; Kaulbach, 1978, pp. 162–170; Moyar, 2008, pp. 287–290; Muchnik, 2009, pp. 100–109; Pieper, 2011, pp. 104–108; Puls, 2013, pp. 104–112; Stolzenberg, 2008, pp. 419–425)

[8] For detailed analyses of the mechanism by which competing incentives are incorporated into a unified determination of the will, see: (Reath, 1989; Timmermann, 2022)

action, as well as its consequences, is therefore imputable to him: He cannot excuse himself by appealing to the external threat, because it was *his* decision to choose between two competing desires and prioritize self-preservation over truthfulness under the given circumstances.

A Kantian response to the contemporary debate over the possible recognition of current LCSs as AMAs thus begins by asking whether they possess practical *selves* of their own, as manifested in the apperceptive capacity to unify desires in light of the unconditional validity of the moral law—a condition that grounds our recognition of ourselves as moral agents: An entity (X) cannot be regarded as a moral agent from a Kantian perspective, which seems to reflect accurately the commonsense notion of moral agency, unless X can be held responsible for its own actions because those actions originate in X's freedom of choice, which ought to be exercised in accordance with X's recognition of the moral law. Freedom of choice consists in choosing among diverse, competing incentives and adopting particular ones as grounds for action, such as respect for the moral law, which demands that one remain truthful under all circumstances. Current LCSs can therefore be recognized as AMAs, I contend, only if they belong to the class of entities possessing practical apperception.

3 Recent Kantian Skepticism in the AMA Debate

Kant's analysis of human agency, grounded in the formal structures of the mind's cognitive and practical activities, invites contemporary philosophers to approach the issue of AI agency through a Kantian lens. A prominent strand of recent Kant scholarship, both immediately preceding and following the public emergence of LLMs, however, has critically responded to the recognition of current AI systems as agents by arguing that they fail to meet the necessary conditions of Kantian subjectivity. I therefore briefly sketch Kantian skepticism about AMAs in order to situate my own assessment within the contemporary debate. The aim of surveying these Kantian approaches is to establish the context for the discussion in Section III, in which I argue that my account offers an advantage over these approaches because it accommodates the epistemic difficulties inherent in the AMA debate.[9]

I take the papers collected in *Kant and Artificial Intelligence*, edited by Kim and Schönecker, to make seminal and systematically well-grounded contributions to the debate over whether LCSs can be recognized as agents, particularly from the standpoint of Kantian ethics. Schönecker develops an argument from moral feeling: On his reconstruction of Kant's model of obligation, finite rational beings cognize the unconditional bindingness of the moral law only through a priori moral *feelings* such as respect and conscience. Since computers lack phenomenal consciousness and hence any feelings, artificial practical reason is impossible. In his view, the decisive difference thus lies not in the transcendental "I think" but in the first-personal "I feel"—and the "I want": Predicates like thinking and wanting must not be applied to beings without an *I* (Schönecker, 2022). Benossi and Bernecker reach a convergent verdict: Even robots with strong AI could not meet the Kantian standards of moral agency—freedom of choice (*Willkür*) and personality—since they follow rules *blindly* rather than acting on their own representations of laws they give to themselves (Benossi & Bernecker, 2022).[10]

Wright approaches the issue of AI agency from the perspective of Kant's political philosophy: Machine

[9] Aside from those centered on practical and moral selfhood, which have direct implications for the AMA debate, there are also approaches to artificial agency inspired by Kant's analysis of cognitive selfhood. Evans explores the possibility of an "apperception engine," as he calls it, designed to model the structure of human apperception through machine learning (Evans, 2022). The "Spatial Apperception Engine," conceived by Soeteman and van Lambalgen, similarly rely on the idea that Kant's transcendental apperception as the source of self-consciousness could be realized within a computational architecture (Soeteman & van Lambalgen, 2024).

Baiasu cautions against premature optimism grounded in a Kantian conception of transcendental apperception by comparing the operations of contemporary AI systems to those of an automatic thermometer, which is neither self-conscious nor a subject, although one might say that it *detects* and *reacts* to room temperature. He circumspectly leaves aside the question of whether they indeed have the capacity to ascribe the results of their operations to themselves (Baiasu, 2022). In this regard, Jo argued that current AI systems are not entitled to recognition as *self*-conscious, even though their capacity to cognize empirical objects can be interpreted as a sign of consciousness that Kant examines in the Transcendental Deduction (Jo, 2025).

[10] The following articles also focus on the Kantian conditions of practical subjectivity, which current AI systems clearly lack: (Chakraborty & Bhuyan, 2023; Manna & Nath, 2021)

autonomy—if there were indeed such a thing—is not identical with human autonomy. LCSs may be programmed to act *rightly*, but they cannot *freely* choose what is right because it is right, nor can they adopt incentives as motivating reasons. The "rightful machine" that she discusses thus remains a non-moral agent, even if its conduct is governed by the priority of right (Wright, 2023). Drawing on a linguistic analysis of AI outputs, Gubelmann supports Kantian skepticism from a different angle, arguing that current transformer-based LLMs are not genuine agents because action requires intentions, and intentions require autonomous goal-setting (Gubelmann, 2024). LLMs thus remain heteronomous mechanisms whose goals are externally imposed, and they therefore cannot perform speech *acts*, although their cognitive self-organization may mark an early step toward future non-biological, genuinely autonomous agents eventually.[11]

[11] These recent objections to recognizing AI as an agent can be understood as reactions to earlier, more optimistic Kantian attempts to treat the Categorical Imperative as a candidate master rule for AI alignment—a trend, in my view, inspired by formal clarity of Kantian ethics, such as: (Powers, 2006; White, 2022)

4 "I think," not "It thinks": Kantian Limits on Assessing Other Minds

These Kantian approaches converge on a common strategy: They identify specific mental capacities as central to Kant's conception of agency and deny that current LCSs possess them. Despite the different aspects of human agency on which their reconstructions focus, the core capacities specified in these assessments can be understood as functions of *practical apperception*, which designates the underlying self-conscious subject to *whom* these capacities are to be ascribed. Schönecker's characterization of the first-personal "I feel" as a capacity that LCSs lack, for example, raises the question of the subject underlying this capacity—namely, who the *I* that feels is. The preceding discussion in Section II has attempted to address this question by drawing a parallel between practical apperception and the cognitive "I think." The conditional claim that, from a Kantian perspective, LCSs cannot be recognized as AMAs unless they possess a capacity that can be characterized as practical apperception is therefore justified, irrespective of the technical question of whether current LCSs *in fact* satisfy this necessary condition.

Within the epistemic limits imposed by Kant's critical philosophy, furthermore, it *cannot* in principle be determined conclusively whether LCSs possess such a mental capacity: In the case of *my* own practical self-consciousness, *I* can directly recognize both diverse desires and their unification into a single determination of the will under the unconditional validity of the moral law. The same holds true of transcendental apperception, whose Deduction is grounded in the fact that sense perceptions are combined into an objective representation in accordance with conceptual rules of synthesis—a mental activity that is accessible only from the *first-personal* perspective of the cognitive subject. I cannot do the same, however, with respect to the mental states of *other* conscious subjects. In his critique of rational psychology, Kant explicitly rejects attempts to attain substantive cognition of the thinking self as it is in itself, regarding them as a misguided form of metaphysics incapable of extending our knowledge (A341–A405; B399–B432). Although Descartes' cogito argument refers only to the existence of a specific *I* that thinks, Kant reads it as exemplifying such an approach to the thinking self *in general* because he construes it as syllogistic reasoning grounded in the unverified universal premise that *everything that thinks exists*.[12] In contrast to the inferential derivation of the thinking subject, Kant's Deduction does not overstep the boundaries of a first-personal perspective: Although *I* can designate the subject underlying *my* mental states by means of the "I think," the same reasoning cannot be extended to formulations such as "she thinks" or "it thinks."

The Kantian assessments considered thus far are therefore vulnerable to Kant's own critique of rational

[12] For Kant's explicit objection to the inference that derives the existence of the thinking self from a universal premise, see: CPR B422 n. However, Descartes himself denies that the cogito is a syllogism, presenting it instead as an immediate intuition. For a discussion of Kant's inaccurate interpretation of Descartes, see: (Beyssade, 2008; Longuenesse, 2008)

psychology insofar as they are grounded in the claim that current LCSs do not genuinely think and therefore cannot be regarded as self-conscious: Kant himself would not license any assertoric judgment about other minds based on my first-personal perspective. Yet these Kantian approaches *as stated* deny self-consciousness to LCSs on the grounds that their operations are fundamentally different from human thought. In keeping with the spirit of Kant's critique of rational psychology, however, the AI-generated outputs, regardless of how closely they resemble those of human consciousness, *neither* confirm *nor* refute the assumption that they are underlain by mental capacities and a subject that possesses them. The burden of proving that AI systems in their current form are not AMAs because they lack phenomenal consciousness, for example, cannot be met, since the question of whether they possess such consciousness cannot be conclusively resolved on the basis of third-personal judgments such as "it thinks" or "it does not think."

The Kantian approach developed in this paper would also face the same criticism if I were to advance the substantive claim that current LCSs do not *in fact* possess any capacity comparable to human practical apperception. Again, I have no reliable access to any inner states LCSs *might* possess. Inherent differences between the human brain and AI's algorithmic architecture are insufficient to substantiate the bold claim that what the latter instantiates does not constitute genuine mental activity comparable to that of human agents.[13] Adhering to Kant's strict epistemic limits thus precludes any *assertoric* judgment concerning other minds. I first delimit the scope of my argument to the conditional claim that *if* LCSs cannot be taken to possess a capacity comparable to practical apperception, they cannot be recognized as Kantian subjects and, accordingly, as AMAs.

[13] The assumption that consciousness is *multiply realizable* across biological and non-biological substrates establishes no more than the in-principle possibility of non-biological consciousness: It does not show that the outputs of any current LCS are generated by conscious processes; such an attribution requires independent, system-specific evidence that the system instantiates the relevant consciousness-supporting functional or causal organization. For an in-depth discussion of the multiple realizability of consciousness in machine architectures, see: (Polger & Shapiro, 2016, pp. 123–149)

III. Withholding Provisional Agent Status from LCSs

To prevent the absurd consequences of bracketing the minds presumed to underlie others' actions, Kant permits the *provisional* recognition of others as subjects based on embodied *signs* of subjectivity (CPR A359–60). The duty to respect other human beings as ends in themselves thus remains binding, even though *I* cannot verify from my first-personal standpoint that they think in the same way *I* do: They could still be philosophical zombies, which merely act and react without phenomenal consciousness. At the same time, Kant carefully avoids overreaching by restricting the implications of such provisional recognition to the modest claim that the supposition of an underlying subject is not *logically contradictory*. It thus remains undetermined, in principle, whether there *is* indeed an underlying subject to which the embodied expressions are to be ascribed. I am nonetheless entitled to treat other human beings as minded entities, despite the theoretical indeterminacy, insofar as empirical *signs* support recognizing them as such and that recognition has not been defeated by conclusive counterevidence. These signs thus serve to protect our shared ethical life against the threat of radical solipsism.

The conditional thesis I have established therefore marks the Kantian limit of what can be established about artificial moral agency: Any judgment about whether a LCS in fact satisfies the necessary condition remains at best provisional and, even if affirmative, would not by itself establish AMA status. At the same time, Kant's strategy of permitting the provisional recognition of others' minds on the basis of a rational inference from embodied signs also prevents my conditional thesis from being practically empty: Although a LCSs' status as a subject—and, accordingly, their eligibility for recognition as an AMA—cannot be conclusively determined without establishing whether it satisfies the necessary condition, *I* can nevertheless rationally determine and justify *my* provisional stance toward these mechanical entities based on the presence or absence of apparent signs of subjectivity. Before closing the discussion, I argue that the recognition of current LCSs as AMAs should be defeasibly *withheld* on the grounds that the abductive reasoning supporting Kant's provisional recognition of human beings as subjects cannot yet be extended to entities whose operations are generally taken to be governed by externally imposed directives. On this assessment, I maintain, current LCSs are therefore more analogous to mere tools than to human agents.

1 Why Don't I Regard Other Human Beings as Philosophical Zombies?

Despite his explicit recognition of other human beings as moral agents with equal capacities and rights, Kant specifies neither the signs of cognitive subjectivity nor those of practical apperception. It thus remains an interpretive task to develop reasonable standards for Kantian recognition of other minds.

Building on the discussion in the previous sections, the following three conditions are regarded as necessary for practical apperception, understood as the foundation of human moral agency: the possession of desires, the capacity to incorporate them into maxims, and the cognition of the moral law. Practical apperception is a subjective faculty for integrating competing desires as incentives into a unified principle of action, namely, into a maxim. The freedom of human beings is manifested in their capacity to take respect for the moral law as a pure rational incentive distinct from the sensuous impulses they share with non-rational beings. Desires thus provide the material for the apperceptive act of a practical subject: I regard myself as a free agent, on Kant's account of freedom, insofar as I can *willingly* resist natural desires through the feeling of respect for the moral law. The recognition of human beings as moral agents thus finds a Kantian justification in their capacity to determine the will in accordance with the moral law, respect for which functions as a pure rational incentive competing with incentives from self-interest, and thereby to exercise freedom responsibly. I *ought* to act morally—leaving aside the metaethical issue of what it means to act morally—and this pure "ought" implies that I *can* indeed act morally, for otherwise the "ought" would be meaningless.[14] Put differently, one is able to be moral only if the possibility of not being moral remains open: I do not morally praise mere tools for being useful to us, despite their selfless service, because they fundamentally lack any capacity to determine themselves *otherwise* than in accordance with a user's commands. The predicate "moral" therefore cannot be attributed to anything that lacks freedom of choice. Kant's argument for freedom of the will adopts the same idea: The *a priori* consciousness that we ought to act morally would be meaningless unless the subject were able to act in accordance with it: A forced act is not regarded as genuinely moral because such an act occurs independently of the agent's free will. Kant thus clarifies the necessary relationship between the moral law and freedom by designating the latter faculty as the *ratio essendi* of the former, which binds me as a rational subject, whereas the moral law constitutes the *ratio cognoscendi* of freedom—namely, the ground through which I become aware of my freedom (CPrR 5:4 n.).

Kant's Categorical Imperative (CI) constitutes the highest norm in the exercise of freedom. In one of its

[14] Although Kant repeatedly invokes the principle that "ought implies can," its precise import and systematic place in his philosophy have remained the subject of persistent debate. For helpful discussions of this principle in Kant's philosophy, see: (Kohl, 2015; Stern, 2004)

well-known formulations articulated in the CPrR, the CI commands us to act only on maxims—namely, the subjective principles of action—whose universal adoption can be rationally willed (CPrR 5:30). The CI thus presupposes the existence of competing maxims: The truthful decision in the preceding illustration given by Kant follows from a strict maxim which permits no lying under any circumstances. The maxim underlying the truthful decision—for example, "I will not tell a lie under any circumstances"—as well as its enactment in particular actions have moral significance only if the maxim is adopted based on pure rational incentives. Suppose, then, that he is asked why he adopts such a maxim. Given that his choice is genuinely moral, he would respond by appealing to a purely rational ground, such as that telling a lie is morally wrong. By contrast, a merely pragmatic justification based on the possible benefits of truthfulness fails to meet the pure rational standard because the incentive to pursue such benefits is self-interested, and therefore *impure* in character. As rational agents, human beings are capable of recognizing the rightful order of the world, which imposes itself on us in the form of an "ought." Such rational recognition gives rise to an incentive to act in accordance with this order, which must compete with other, non-rational incentives in determining the will. Kant's CI thus commands me to act on a maxim grounded in a pure rational incentive, the universal adoption of which all rational agents could will. His argument describing the moral law as the *ratio cognoscendi* of freedom thus rests on the recognition of complete leeway in choosing which incentive to follow—that I have designated as *incorporating* incentives.[15]

The ascription of practical apperception to other human beings is thus justified by an *abductive* argument: Moral conduct observable in experience provides us grounds for positing an underlying subject as its best explanation. If there were no subject who recognizes the moral law and thereby experiences respect for it as a pure rational incentive, such conduct would remain unexplained. The abductive character of Kant's reasoning manifests itself in his critical assessment of the metaphysical idea of freedom in the CPR (CPR A444/B472–A447/B475): The affirmation of freedom conflicts with the causal completeness of nature. To suppose an entity whose actions are not governed by causal laws is to posit an exception to the universal causal order. Thoroughgoing causal determination without freedom, however, leads to an infinite regress, since every event is conditioned by a preceding cause that is itself the effect of a still earlier cause. Kant's appeal to the moral law as the *ratio*

[15] Kant's theory of radical evil in *Religion* also rests on agential freedom in the choice of incentives. Original sin, in his rationalist reading of the Bible, is understood as a crooked propensity to subordinate rational incentives to non-rational ones, even while recognizing that one ought not to do so (Religion 6:29–37).

In using the term "incorporate," I follow Allison's influential interpretation, according to which human freedom lies not in choosing which desires arise but in deciding which of them to adopt as incentives; desires themselves are not products of freedom but are merely given. In his seminal study of Kant's *Religion*, Allison distinguishes two levels of agential activity: an initial, at which the agent experiences incentives, and a deeper level, at which the agent decides which incentives to incorporate into a maxim and which to set aside. The Kantian freedom of choice, on Allison's interpretation, operates at the second level (Allison, 1990, pp. 30–53).

cognoscendi of freedom thus provides a practical basis for affirming freedom despite the theoretical impasse concerning its place within the universal causal order: Given that I am conscious of the moral law as unconditionally binding—a consciousness Kant designates as a *fact* of reason—I must regard myself as free; otherwise this consciousness would remain unintelligible. Kant's appeal to the essential intelligibility of the world does not contradict his theoretical agnosticism about freedom: I am compelled to affirm that at least one free being exists in the universe—namely, myself—without thereby resolving the ontological question of whether I am *in fact* free. The conclusion is nonetheless rationally justified insofar as it is drawn from the inherent need of reason for explanation. I am entitled to regard myself as free insofar as my awareness of the moral law as unconditionally binding is better explained by supposing that I am free than by any alternative explanation.

Kant's recognition of animal minds by analogy suggests the possibility of extending comparable recognition to mechanical other minds: It is rationally justified to infer from animal behavior, such as dam-building by beavers, that animals possess the capacity to act on mental representations. The principle justifying such inference "on the basis of their actions" lies in the fact that "we have the same ground for counting animals, with respect to the determination in question, as members of the same genus with human beings, as humans" (CJ 5:464 n.). The common genus, on which the provisional recognition of animal minds is grounded, seems to be that of *living beings*. Kant understands life as "the faculty of a being to act in accordance with laws of the faculty of desire" (CPrR 5:9 n.). The activities of living beings, including my own, are thus expressions of desires, which non-living beings lack:[16] From my first-personal perspective, it is certain that I feel desires. Insofar as some of my desires arise from bodily conditions, such as hunger from the empty stomach, I can reasonably attribute comparable mental states to other living beings that share similar body structures.[17] Kant's comparison of beavers' building activity with human construction, which is guided by mental representations, is therefore justified insofar as supposing that beavers have a representational capacity—although one less sophisticated and versatile than human thought—provides the best explanation of their behavior, which therefore may be understood as an embodied sign of mental activity.

[16] One might argue directly against recognizing LCSs as AMAs on the grounds that they are not living beings and thereby lack desires, the possession of which constitutes a necessary condition for practical apperception. I do not pursue this line of argument, however, because it seems to rely on a stipulative definition of life when advanced without sustained engagement with Kant's teleological conception of living organisms: I believe that Kant's philosophy of life offers conceptual resources for a promising approach to artificial moral agency, although I do not further develop these resources in this paper.

[17] It bears emphasizing once again that the recognition of animal minds is also *provisional*, given Kant's strict standards for philosophical cognition. The inference supporting such recognition is defeasible, although it may be rationally well grounded. Nevertheless, its rational warrant justifies *my* treatment of other embodied beings, such as animals, as minded.

Kant's recognition of other minds through abductive reasoning thus rests on embodied structures of desire shared by living beings. Human beings are distinguished by their capacity to refrain from blindly pursuing their desires out of respect for the moral law, arising from rational cognition of the law. *I* do therefore regard other human beings not as philosophical zombies but as free agents equal in status to *myself*, insofar as their actions are best explained by positing mental states of their own: They may thus be *provisionally* regarded as satisfying the three conditions of practical apperception. The duty to treat other human beings with respect thus remains binding insofar as the grounds for such recognition are not undermined by relevant counterevidence. Withholding moral-agential status from non-human animals is likewise abductively justified insofar as their behavior can be adequately explained without positing a higher-order capacity for rational self-control constitutive of freedom as the *ratio essendi* of the moral law in the Kantian sense.

Withholding agential status from current LCSs on Kantian grounds is therefore justified if their operations can be well explained without positing practical apperception. Leaving aside the first and third conditions of practical apperception, which would require in-depth engagement with debates concerning desire and cognition in contemporary philosophy of mind, I focus on its second condition—namely, the capacity to incorporate distinct desires into a unified determination of the will. If current LCSs are provisionally judged not to satisfy this *necessary* Kantian condition, I am not entitled to regard them as subjects, and accordingly, as AMAs.

To the extent that their behavior is normatively constrained, current LCSs are designed and trained to approximate conformity with standards specified by human designers and institutions, rather than standards they autonomously legislate for themselves.[18] On this account, they do not themselves choose the incentives that guide their operations—leaving aside the question of whether they experience incentives at all—whereas the second condition of practical apperception concerns the free activity of choosing among competing incentives and incorporating selected ones into maxims.[19] I thus have a straightforward explanation for their seemingly moral responses: Their operations are constrained by *programmed* ethical guidelines. The abductive reasoning that supports recognizing other human beings as minded subjects cannot therefore be extended to these mechanical entities, whose operations can be adequately explained without positing an underlying subject: In his argument for freedom, as has been discussed, Kant takes our consciousness of the moral law's unconditional validity as a fact of reason, without which many human actions would remain unexplained. Being a moral agent implies freedom. Human beings satisfy this condition insofar as they are capable of choosing whether to conform to or violate the law. Their embodied actions are taken to signify underlying acts of free choice insofar as these actions are best explained by positing such acts. By contrast, current LCSs *need not* be regarded as free agents, since their operations remain intelligible without positing a consciousness of their own that chooses among desires and *incorporates* some of them into maxims—just as I need not suppose that a machine soul dwells within my laptop

[18] OpenAI's technical report illustrates how contemporary machine-learning systems are governed by normative rules designed by their developers (OpenAI, 2024). In his paper discussed above, Gubelmann aptly points out the absence of autonomous will in current AI models, which are therefore constrained to conform to externally given moral standards (Gubelmann, 2024). The contemporary debate over AI alignment rests on the requirement that LCSs' operation can—and must—be aligned with human moral standard rather than standards that these systems autonomously choose for themselves.

Whether the results generated by current LCSs are expressions of mere heteronomy is in fact a question that requires in-depth investigation—which I unfortunately cannot undertake here because of space limitations. In the absence of further discussion over AI autonomy, I contend that the burden of proof lies with those who argue that LCSs may operate autonomously, contrary to the prevailing understanding of their operation.

[19] If they do not possess desires at all—the possession of desires constituting the first condition of practical apperception—then the failure to satisfy the second condition might follow trivially.

and translates my keystrokes into letters, because I have a better explanation of how those keystrokes are converted into written signs through computational processing. Likewise, our knowledge that current LCSs are governed by preset guidelines that they cannot themselves *choose* provides sufficient justification for my provisional withholding of agential status from them.[20] Optimistic assessments, which occupy an important position in the recent debate over artificial moral agency, can be understood as attempts to separate agency from the faculty of choice, given that current LCSs are generally regarded as lacking the latter. In light of Kant's grounding of practical subjectivity in practical apperception, however, agency without the faculty of choice is precluded, since behaviors cannot be regarded as agential unless they result from the subject's free incorporation of competing desires into a maxim.[21]

[20] I need not answer detailed questions concerning the exact mechanisms governing LCSs' ethical behavior, such as which guidelines control AI-generated outputs and how they do so, to justify my *provisional* stance, given that I can appeal to reliable sources explaining those mechanisms—just as I need not fully understand the computational mechanism by which keystrokes are converted into written characters in order to dismiss fantastical explanations. The fact that the stance is defeasible and may be revised in light of counterevidence that undermines its justification does not itself undermine the rationality of adopting it: We need not be climate scientists to maintain that climate change is real.

[21] The *prima facie* tautological claim that LCSs can be genuine agents only if they can freely choose reveals its significance when set against the optimistic trend in the AMA debate. Wright is therefore right to observe that LCSs cannot be autonomous in the same sense in which human beings understand themselves to be autonomous; the assertoric force of her assessment should, however, be moderated if it is to remain consistent with the Kantian framework, which recognizes that direct access to other minds is unavailable.

IV. In Lieu of Concluding Remarks: The Deceptive Appearance of AI's First-Personal Responses

Instead of recapitulating the major findings of the previous discussion—which I hope are sufficiently clear to require no reiteration—I devote these concluding remarks to a Kantian proposal for preventing unnecessary confusion that result in the unwarranted anthropomorphization of LCSs:[22] AI-generated outputs, especially responses to written prompts, should not be formulated using the first-personal pronoun "I," because the systems generating them cannot presently be taken to possess a self that corresponds to that expression. The use of first-personal expressions has moral implications that are by no means trivial: I ought to respect other human beings as ends in themselves insofar as they have selves equal to myself. Allowing LCSs to formulate their responses in first-person sentences may thus have a disguising effect, making it appear as though they *act* in the same way as human beings, although their operations cannot presently taken as analogous to human conduct. Although submarines may be said to *swim*, they do not swim in the same sense that human beings do.[23] By the same token, current LCSs cannot, even provisionally, be said to *act* in the original sense of the term and therefore cannot be regarded as AMAs. AI's use of first-personal expressions is thus unnecessarily misleading.

[22] Emotional reactions to the retirement of GPT-4o from ChatGPT illustrate that current LCSs are gaining forms of recognition comparable to those accorded to human or animal companions. However heartfelt they may be, these reactions are philosophically misguided insofar as they presuppose that users were accompanied by a self-like entity whose disappearance constitutes a genuine interpersonal loss. For an analysis of mourning reactions to the retirement of GPT-4o, see: (Donard & Ribeiro, 2026).

[23] I borrow this analogy from Schönecker's article (Schönecker, 2022, pp. 185–186).

References

Allison, H. (1990). *Kant's Theory of Freedom*. Cambridge University Press.

Baiasu, S. (2022). The Challenge of (Self-)Consciousness: Kant, Artificial Intelligence and Sense-Making. In H. Kim & D. Schönecker (Eds.), *Kant and Artificial Intelligence* (pp. 105–128). Walter de Gruyter GmbH.

Benossi, L., & Bernecker, S. (2022). A Kantian Perspective on Robot Ethics. In H. Kim & D. Schönecker (Eds.), *Kant and Artificial Intelligence* (pp. 147–168). Walter de Gruyter GmbH.

Beyssade, J.-M. (2008). Descartes' "I Am a Thing That Thinks" versus Kant's "I Think". In D. Garber & B. Longuenesse (Eds.), *Kant and the Early Moderns* (pp. 32–40). Princeton University Press.

Chakraborty, A., & Bhuyan, N. (2023). Can Artificial Intelligence Be a Kantian Moral Agent? On Moral Autonomy of AI System. *AI and Ethics*, *4*, 325–331.

Coeckelbergh, M. (2009). Virtual Moral Agency, Virtual Moral Responsibility: On the Moral Significance of the Appearance, Perception, and Performance of Artificial Agents. *AI and Society*, *24*, 181–189.

Coeckelbergh, M. (2010). Robot rights? Towards a Social-Relational Justification of Moral Consideration. *Ethics and Information Technology*, *12*(3), 209–221.

Donard, V., & Ribeiro, J. C. (2026). Technological mourning after AI updates: mental health and well-being in the GPT-4o/GPT-5 transition. *Academia Mental Health and Well-Being*, *3*(2). https://doi.org/10.20935/MHealthWellB8256

Evans, R. (2022). The Apperception Engine. In H. Kim & D. Schönecker (Eds.), *Kant and Artificial Intelligence* (pp. 39–103). Walter de Gruyter GmbH.

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

Grüne, S. (2009). *Blinde Anschauung : die Rolle von Begriffen in Kants Theorie sinnlicher Synthesis*. Klostermann.

Gubelmann, R. (2024). Large Language Models, Agency, and Why Speech Acts are Beyond Them (For Now) – A Kantian-Cum-Pragmatist Case. *Philosophy & Technology*, *37*(32), 1–24.

Gunkel, D. (2018). The Other Question: Can and Should Robots Have Rights? *Ethics Inf Technol*, *20*, 87–99.

Hoffmann, T. S. (2002). Gewissen als praktische Apperzeption. Zur Lehre vom Gewissen in Kants Ethik-Vorlesungen. *Kant-Studien*, *93*(4), 424–443.

Jo, S.-Y. (2025). A Kantian Approach to the Possibility of Mechanical Consciousness. *Lessico di Etica pubblica*, *2*, 67–76.

Jo, S.-Y. (2026). I Think and I Will: Revisiting Practical Apperception as the Kantian Foundation of Personality. *CHUL HAK SA SANG : Journal of Philosophical Ideas*, *100*, 33–66.

Kant, I. (1996). Religion within the boundaries of mere reason (G. di Giovanni, Trans.). In A. W. Wood and G. di Giovanni (Eds.), *Religion and rational theology* (pp. 39–216). Cambridge University Press.

Kant, I. (1998). *Critique of pure reason* (P. Guyer & A. W. Wood, Trans. & Eds.). Cambridge University Press.

Kant, I. (2000). *Critique of the power of judgment* (P. Guyer, Ed. & Trans.; E. Matthews, Trans.). Cambridge University Press.

Kant, I. (2015). *Critique of practical reason* (M. Gregor, Trans.; 2nd ed.). Cambridge University Press.

Kaulbach, F. (1978). *Das Prinzip Handlung in der Philosophie Kants*. Walter de Gruyter.

Kohl, M. (2015). Kant and 'ought implies can'. *The Philosophical Quarterly*, *65*(261), 690–710.

Longuenesse, B. (1998). *Kant and the Capacity to Judge*. Princeton University Press.

Longuenesse, B. (2008). Kant's "I Think" versus Descartes' "I Am a Thing That Thinks". In D. Garber & B. Longuenesse (Eds.), *Kant and the Early Moderns* (pp. 9–31). Princeton University Press.

Manna, R., & Nath, R. (2021). Kantian Moral Agency and the Ethics of Artificial Intelligence. *Problemos 100*, 139–151.

Moor, J. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, *21*, 18–21.

Moyar, D. (2008). Unstable Autonomy: Conscience and Judgment in Kant's Moral Philosophy. *Journal of Moral Philosophy 5*, 327–360.

Muchnik, P. (2009). *Kant's Theory of Evil*. Rowman&Littlefield.

OpenAI. (2024). *GPT-4 Technical Report* https://doi.org/10.48550/arXiv.2303.08774

Pieper, A. (2011). Zweites Hauptstück (57-71). In O. Höffe (Ed.), *Immanuel Kant: Kritik der praktischen Vernunft* (pp. 101–116). Akademie Verlag.

Polger, T. W., & Shapiro, L. A. (2016). *The Multiple Realization Book*. Oxford University Press UK.

Powers, T. (2006). Prospects for a Kantian Machine. *IEEE Intelligent Systems*, *21*(4), 46–51.

Puls, H. (2013). *Funktionen der Freiheit*. Walter de Gruyter GmbH.

Reath, A. (1989). Kant's Theory of Moral Sensibility: Respect for the Moral Law and the Influence of Inclination. *Kant-Studien*, *80*, 284–302.

Schönecker, D. (2022). Kant's Argument from Moral Feelings: Why Practical Reason Cannot Be Artificial. In H. Kim & D. Schönecker (Eds.), *Kant and Artificial Intelligence* (pp. 169–188). Walter de Gruyter GmbH.

Soeteman, A., & van Lambalgen, M. (2024). Spatial Unity for the Apperception Engine. *International Journal of Approximate Reasoning*, *165*, 1–22.

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62–77.

Stern, R. (2004). Does 'Ought' Imply 'Can'? And Did Kant Think It Does? *Utilitas*, *16*(1), 42–61.

Stolzenberg, J. (2008). The Pure “I Will” Must Be Able to Accompany All of My Desires: The Problem of a Deduction of the Categories of Freedom in Kant’s Critique of Practical Reason. In *Recht und Frieden in der Philosophie Kants : Akten des X. Internationalen Kant-Kongresses* (pp. 415–425). Walter De Gruyter.

Timmermann, J. (2022). *Kant's Will at the Crossroads: An Essay on the Failings of Practical Rationality*. Oxford University Press.

White, J. (2022). Autonomous Reboot: Kant, the Categorical Imperative, and Contemporary Challenges for Machine Ethicists. *AI & SOCIETY*, *37*, 661–673.

Wright, A. (2023). A Kantian Course Correction for Machine Ethics. In *Technology Ethics* (pp. 141–151). Routledge.