



Kantian deontology for AI: alignment without moral agency

Oluwaseun Damilola Sanwoolu¹

Received: 18 February 2025 / Accepted: 20 June 2025 / Published online: 11 July 2025
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2025

Abstract

This paper explores the potential application of Kant's moral philosophy to artificial intelligence (AI) and addresses two major objections. The first objection is that AI cannot fulfill Kant's standards for moral agency. I contend, however, that AI alignment with Kantian principles does not require moral agency in Kant's sense. I propose that the Categorical Imperative (CI) can serve as a useful framework for AI alignment, guiding the creation of maxims governing AI actions and testing their universalizability, particularly using the first principle of the CI which is the formula of the universal law (FUL). The second objection I address is the particularist critique to Kantian universalism, which is that Kantian universalism cannot tell us how to form maxims in a way that it allows sensitivity to context. I maintain that Kant's framework can indeed accommodate context-sensitivity through practical judgment. But since AI are not the kinds of things to have practical judgment, I show that they have a functionally equivalent mechanism—transformer models—which can allow them form maxims that consider morally salient facts. Thus, supporting the claim that AI alignment is possible within a Kantian framework.

Keywords Artificial intelligence · Alignment · Kantian deontology · Formula of the universal law (FUL) · Categorical imperative (CI) · Transformer models

1 Introduction

The field of Artificial Intelligence (AI) ethics is actively debating whether AI systems should be treated as moral agents, capable of making their own ethical decisions [1, 27, 37, 42]. This area of inquiry is often called Artificial Morality, which explores how AI systems can be equipped with moral capabilities, rather than simply reflecting the values of their creators or users [1]. A key question within Artificial Morality is: if AI systems are to be considered moral agents, what ethical principles should guide their actions? This question is closely related to the broader issue of AI alignment, which examines how to ensure that AI systems operate in accordance with human values. Researchers are exploring this challenge by drawing on various ethical theories and frameworks.

One theory is the Kantian ethical framework which is my focus in this paper. Thinking about the application of Kant's

moral philosophy to AI is useful since it raises questions about the connection between kinds of agents (their actual and potential capacities, their relationships to us, etc.) and the appropriateness of applying competing moral frameworks to their situations. What would it mean, for example, to apply Kant's moral philosophy in a context where the agents in question are not moral agents of the kind that Kant would have recognized? Is this a coherent possibility given Kant's moral philosophy, where the idea of the autonomous rational agent is so central to his conception of ethical action?

While a full introductory account of Kantian deontology is beyond the scope of the present paper, I will briefly rehearse Kant's position in Section Three to touch on one of the crucial concepts in his moral philosophy—the Categorical Imperative (CI). I take for granted that readers are familiar with Kant's ideas on duties, rational autonomy and the good will.

Scholars such as Chaly, Mougan & Brand, and Powers have argued for the compatibility between the Kantian ethical framework and machine ethics [8, 26, 28]. However, these approaches are vulnerable to an obvious criticism; AI artifacts do not have the capacity to be the kinds of moral

✉ Oluwaseun Damilola Sanwoolu
oluwaseun.sanwoolu@ku.edu

¹ University of Kansas, Lawrence, USA

agents that Kant had in mind in his moral Philosophy. However, this criticism is too quick, and in this paper, I argue that the fact that these agents are different in kind from Kant's conception of the moral agent does not mean that we cannot align AI with the Kantian ethical framework. In fact, I will argue that AI can be aligned to the Kantian ethical framework without taking the stand that they are the kinds of moral agents Kant had in mind.

Some philosophers have argued that we ought not align AI systems with any ethical theory as there is an absence of moral agreement. Since there are numerous ethical positions, how do we decide which ethical theory to turn to? In this paper, I address this. I reject the idea that the multiplicity of ethical theories provides sufficient justification for bypassing ethical theories as candidates for AI alignment. Instead, I propose that we can turn to ethical theories for AI alignment, specifically Kantian deontology. I maintain that the Categorical Imperative (CI) can be a framework for generating maxims governing the actions of AI, and testing their universalizability, particularly the first formula of the CI.

In defending this view, I address two key objections. The first objection is the claim that AI are not the kinds of things that can be Kantian agents. I respond to this objection even though I accept the truth of the claim that AI is not capable of Kantian moral agency. However, I provide reasons why we do not need to commit to them as Kantian agents to believe that they can be aligned to his ethical framework.

The second major objection is the particularist challenge to Kantian universalism. To address this, we must demonstrate that Kant's framework accommodates moral salience in maxim formation. Drawing from Bremner and Dunn [13, 40] I show how we can turn to practical judgment in human agents to deal with this challenge. If we accept these interpretations of Kant's philosophy, then we can extend this to AI. We can examine if AI has a functional equivalent to practical judgment, which allows them simulate context-sensitivity while adhering to Kantian principles. Thus, my proposal can be established; AI can use the CI as a framework for determining what actions to carry out. Ultimately, my view is that ethical theories, and specifically Kantian deontology, can provide a coherent basis for AI alignment.

I break the rest of the paper into three parts. In the first part, I discuss AI alignment—what it is and approaches to it. In the second part, I consider Kant's deontology to address the restrictions within his theory for the purpose of AI alignment. In the third part, I propose my approach to aligning AI with the Kantian ethical framework and consider possible objections.

2 AI alignment

The AI alignment debates have their roots in discussions outside academic philosophy, mostly among technologists, journalists, and industry experts [20, 34, 35]. The main objective of AI alignment 'is to ensure that powerful AI is properly aligned with human values' [14]. And as Gabriel [14] noted, 'values' here can stand as a placeholder for many things. Although there is no consensus, broadly speaking, the term refers to designing AI systems to address social and environmental concerns without causing undue harm or aggravating current challenges [9, 14, 33]. One well-known illustration of the need for alignment is Bostrom's paperclip maximizer thought experiment [8]. Suppose an AI system is tasked with maximizing paperclip production. Left unchecked, it could convert not only Earth but increasingly large portions of the observable universe into paperclips. While this outcome would technically fulfill the given objective, it clearly contradicts human values. This example underscores a crucial point: certain AI behaviors are universally undesirable, which makes alignment not just a theoretical concern but a practical necessity. Since I am examining AI alignment from a moral philosophical perspective, I use 'values' to refer specifically to moral values.

Scholars such as Cecchini, Gabriel and Morley [9, 14, 28] have noted that there are two main challenges in AI alignment: the technical and the normative. The technical part of this challenge seeks to answer the *how* question—how can we formally encode values or principles in AI systems to ensure they act as they should [14]? The normative challenge, on the other hand, is concerned with the *what* question—what ethical principles or frameworks should we align AI systems with?

A key question is how these two challenges relate to one another. Are they independent, or do they influence each other? One perspective, known as the *simple thesis* [14], holds that we can address the technical aspect of AI alignment independently from the normative aspect, allowing any value system to be encoded later. Another view argues that the two aspects are interdependent. One version of this view suggests that the technical methods used to build AI systems determine which values can be encoded, while the opposite version holds that knowing what values we want to align AI with should inform how we design the systems to align with those values.¹

So far, in addressing the second part of the alignment challenge, there have been two broad paths to engage in this discussion. The first path has been with respect to Artificial morality, and the second path has been to altogether bypass the discussion of Artificial morality. Concerning the

¹ For more discussion on the relationship between the technical part and the normative part of the alignment problem, see Gabriel (2020).

first path, there have been three major approaches. These approaches are: Top-Down, Bottom-Up, and Hybrid [1]. Top-Down approaches are basically rules-based approaches to machine morality. Theories that have been included in this group are, for example, The Golden Rule, Isaac Asimov's four laws of robotics, Utilitarianism, Kantian deontology etc., [1].

Bottom-Up approaches reject rule-based approaches, as they do not believe that we can settle on a specific moral theory. Bottom-up approaches rest instead on developing moral sensibilities within the AI entities in a process of moral education in order to have the artifact make decisions in accordance with morally good principles. This is often likened to the process of moral socialization we see in human childhood.

Hybrid approaches combine elements of both top-down and bottom-up methods, aiming to overcome the limitations inherent in each. As a result, hybrid approaches are faced with a different kind of challenge. As Collin et. al [1] states, they must 'mesh both diverse philosophies and dissimilar architectures'. This means building a system that is not only able to apply ethical principles, but one that also has an appreciation for context.²

Moreover, in addressing the normative aspect of the AI alignment challenge through the second approach—bypassing artificial morality—we observe scholars stepping outside ethical theories. Primarily, this is because of the concern over which specific ethical theory we ought to align AI with given the wide variety of available theories. Gabriel [14] framed it this way; 'there isn't any singular ethical theory that captures the entire truth about issues concerning morality'. While this is true, the implication of such view is that we should ignore ethical theories when it comes to discussions around AI alignment.

However, I maintain that the lack of consensus on the correct ethical theory is not sufficient grounds to bypass ethical theories in this discussion. Concerning human morality, for example, one could adopt pluralist ethical approaches. Although ethical theories all have strengths and limitations, yet we continue to seek moral guidance despite the lack of agreement on a single 'correct' theory. And even if there is the 'right theory', a lack of consensus about such a theory doesn't mean we ignore it or others in the quest of seeking moral guidance. Nevertheless, I propose that we should still take ethical theories into account when discussing AI alignment. As previously mentioned, I will focus on Kantian deontology in this context.

3 Kantian deontology, AI alignment and objections

For Kant, determining if an action is right or wrong depends on if such action conforms to the Categorical Imperative (CI). The CI is usually presented in terms of five formulas. The first being the Formula of universal law (FUL), which states that 'act only according to that maxim whereby you can at the same time will that it should become a universal law' [24]. The second formula is the formula of the universal law of nature which states 'act as if the maxim of your action were to become through your will a universal law of nature' (*Ibid*). The third formula is the formula of the end in itself which is presented as 'act in such a way that you treat humanity, whether in your own person or in the person of another, always at the same time as an end and never simply as a means' [24]. This version is closely linked to human reason, as Kant believed that rational beings exist as ends in themselves. The fourth formula is the formula of autonomy (FA) 'the Idea of the will of every rational being as *a will that legislates universal law*' [24].³ The last formula is the formula of the kingdom of ends which is 'every rational being must so act as if he were through his maxim always a legislating member in the universal kingdom of ends' [24].⁴

In this paper, my focus is primarily on the first formula of the CI which is the FUL. This is partly because the other formulas impose strict requirements tied to rationality, the will and moral agency, which, as will be discussed later, present challenges for non-human agents. Since Kant thought of the different formulas of the CI as one and the same, and we accept the interpretations of scholars like Allison, Guyer and Timmermann [2, 17, 39] then the applications of the different formulas would produce the same result. Using the FUL thus offers a promising avenue for AI alignment since it can produce the same result as the other formulas of the CI without the restrictions on the will, rational humanity etc. Therefore, aligning AI systems with the principle that they should act only in ways that could be universally accepted as a law could be a strong contender when determining the appropriate principles for AI alignment.

3.1 Aligning AI with the Kantian ethical framework: what has been done so far

One of the first studies concerning artificial morality and the Kantian ethical framework has been done by Thomas

² See Allen et. al (2005) for more discussion on Top-down, bottom-up and developmental approaches, and Hybrid approaches to artificial morality.

³ The fourth version of the CI is tied to the autonomy of the will of persons, and it sums up why the first two principles of the CI are categorical Imperatives and not Hypothetical imperatives.

⁴ This formula follows directly from the FA and it is considered as the one that combines the earlier four formulas (Geiger, 2015; Johnson & Cureton, 2022).

Powers. Powers examined how a machine could possess artificial morality within the Kantian ethical system [31]. In doing this, he considered how artificial morality can fit within this modified account and what challenges this endeavor might face. Although the debate concerning AI alignment had not fully developed at the time of Powers' work, we can view his work as one of the earliest attempts at aligning AI with Kantian deontology.

Powers suggested how machines can function as moral agents within the Kantian ethical framework [31], however I maintain that we will face certain limitations if we do so.⁵ Considering Moor's hierarchical classification of moral agents, we can say that the kinds of moral agents Powers had in mind are explicit moral agents. Moor classified moral agents into four groups: ethical-impact agents, Implicit ethical agents, explicit ethical agents and full ethical agents. For him, ethical-impact agents are as their name implies; they are those kinds of technology that have an ethical impact in the world. Implicit ethical agents are machines or software systems designed with a focus on safety or critical reliability considerations. They are programmed in a way that inherently promotes ethical behavior. Explicit ethical agents are AI systems capable of explicitly representing ethical principles and effectively using this knowledge to guide their behavior.⁶ Lastly, a 'full ethical agent can make explicit ethical judgements and generally is competent to reasonably justify them' [27].

Considering these, Powers recognizes that machines cannot be full ethical agents in the Kantian sense. Since explicit ethical agents are systems that can detect and evaluate morally relevant information to guide their behavior, his proposal fits neatly into this category. Thus, in aligning AI systems to Kantian deontology, Powers considered AI systems as explicit ethical agents. But should we consider AI as moral agents to align AI to Kantian deontology?

Only full moral agents, as outlined in Moor's hierarchy, fit within the framework of Kantian deontology.⁷ This is because Kant's deontology is not solely about adhering to rules; it encompasses concepts such as freewill, good will and rational autonomy, that AI artifacts cannot possess. Regarding rational autonomy, if we were to encode Kantian duties or laws into an AI system, it would be compelled to follow these laws, as that is how it is designed to operate. This circumstance contradicts the principle of rational

autonomy. Typically, machine outputs are probabilistic, meaning that they are heteronomous in nature rather than autonomous; their source of law is from external programming rather than self-determination. In contrast, a rationally autonomous person acts according to their own will and not that of others [22]. This distinction illustrates why AI cannot be aligned with the fourth version of the CI. According to Kant, rational agents are not beings who are merely instructed on what to do; they must through reason generate maxims, apply the CI and fulfill their duties based on that reasoning.

Kantian moral agency also presupposes free will, which Kant connects to the ideals of moral perfection and rational autonomy [7]. In simple terms, possessing free will means that an agent can act according to their own will, free from external constraints that would otherwise dictate their actions. This freedom is essential because, for Kant, true moral worth arises when an agent chooses to act according to the moral law out of respect for that law, as it must be an act that stems from the agent's own autonomous will rather than from coercion, external influence, or in this case, internal programming. AI systems, however, lack this critical element of free will; although they may not require mediation at every step of the decision-making process as they can be probabilistic in their outputs, they still operate based on pre-coded algorithmic rules. This operational independence is not equivalent to Kantian autonomy, as it lacks the capacity for self-legislated moral reasoning. The AI's decisions are, at their core, the product of encoded rules and data patterns rather than genuine choices made from a self-governing will.

Furthermore, free will in the Kantian sense involves an interplay between negative and positive freedom.⁸ Negative freedom entails the possibility of choosing otherwise—the ability to deviate from predetermined pathways based on independent judgment. Positive freedom, by contrast, refers to the ability to act according to one's own rational will specifically, in accordance with moral law. While negative freedom highlights the agent's independence from causal determination, positive freedom emphasizes rational self-legislation. Human agents, in Kant's framework, possess both: they can act independently of empirical causes and, ideally, govern their actions through rational moral principles. AI systems, however, lack both aspects. They are constrained by pre-programmed architectures, optimization objectives, and statistical learning from data. Consequently, they do not exhibit negative freedom, as they cannot truly deviate from causal determination, nor do they exhibit positive freedom, as they cannot will or legislate moral law from

⁵ I discuss some of these limitations on the next page.

⁶ See Moor (2011, 18), for examples of these kinds of agents according to his classification.

⁷ Several scholars such as Rawls (2000), Deligiorgi (2012), Manna & Nath (2021), Johnson & Cureton (2022) have argued that Kant's moral agents are rational human agents who Kant considers as not acting according to inclinations but according to what the principles of pure reason require. They have grounded this on Kant's discussions on concepts such as autonomy, freedom, and goodwill.

⁸ I am thankful to an anonymous reviewer for highlighting the distinction between negative freedom of the will and positive freedom of the will.

reason. Since AI operates within the constraints of its programming and cannot independently pursue moral principles, it fundamentally lacks the autonomy required for true moral agency according to Kantian ethics.

Given these limitations, I argue that we refrain from ascribing moral agency to AI, particularly within the Kantian ethical framework.⁹ This leads to a critical question: can ethical principles be applied to actions without recognizing the actors as moral agents?

To address this, consider how we teach moral principles to children. For example, in teaching a child the value of honesty, we often begin by modeling truthful behavior, demonstrating that honesty is preferable even when inconvenient. A child might imitate this behavior, choosing truthfulness despite potential discomfort. However, even though the child performs a morally sound action, we would not consider them a fully developed moral agent. They are mimicking learned behavior rather than acting from an internalized understanding of ethical principles.

Similarly, a child might adhere to a parent's ethical precepts without comprehending the underlying framework. For instance, a parent might forbid lying based on their own moral code, and the child, through obedience, consistently tells the truth. While the child's actions align with an ethical principle and produce a morally good outcome, they lack the autonomous understanding necessary for moral agency in the Kantian ethical framework.¹⁰

These scenarios illustrate a crucial point: a morally good action can occur without full moral agency. This distinction is particularly relevant when considering artificial intelligence. AI systems can be designed to act in accordance with moral guidelines without possessing the capacity for independent moral reasoning.

Some might argue that because AI's actions have moral consequences, we should consider them moral agents. In response, I ask: Is it necessary to view all agents whose actions have moral consequences as moral agents? In answering this, let's first note that we would be wrong to assume that all entities or systems whose actions have normative or moral outcomes are moral agents.

Consider industrial software systems. These systems are designed to function according to specific instructions, and

we evaluate their success based on how well they follow these pre-set guidelines. A software system can be said to have performed correctly or incorrectly (a normative judgment) based on its adherence to those specifications, but we wouldn't necessarily call its actions moral or immoral.

Moreover, if a banking software accidentally overpays someone, it has moral consequences, yet we don't attribute moral judgment or agency to the software itself. Similarly, if a sniffer dog fails to detect illegal drugs, its failure may lead to serious moral outcomes, but we still wouldn't regard the dog's actions as worthy of praise or blame. These are examples of entities whose actions are judged based on how well they perform specific tasks but cannot be considered moral agents.

Second, consider entities without reasoning, like a rock, which can be involved in actions with moral implications. If a rock falls on someone and causes their death, we wouldn't say that the rock acted immorally. There's no way to evaluate whether the rock acted in line with any kind of intention or specification. Yet, the death itself has moral and legal significance. This shows that the existence of moral consequences alone does not justify calling the entities behind them a moral agent.

In contrast, if we consider a grocer who has habitually refrained from inflating prices of goods over the years, even if this behavior has become automatic, we still recognize it as a moral action. This is because morality is contingent on the nature of the entity performing the action. Humans, unlike AI or non-living objects, possess the capacity for moral deliberation and autonomy. Therefore, the fact that AI systems may generate actions with moral consequences is not enough to classify them as moral agents. And we can still take the consequences of their actions seriously without ascribing moral agency to them.

Indeed, some scholars such as Talbot et. al, [38] argue that AI systems should not be held to deontological standards at all, on the grounds that deontology, unlike consequentialism, requires moral agency. Since AI lacks core features of moral agency (such as autonomy, moral responsibility, or the capacity for moral motivation), they conclude that AI cannot be subject to deontic evaluation. I agree with the diagnosis that AI systems are not moral agents, and so cannot be assessed in terms of duty, praise, or blame in the same way humans can. However, I reject the prescription that follows namely, that this rules out the application of deontological principles to AI design and behavior.

My proposal hereafter reframes the issue: while AI systems cannot be Kantian agents, they can still be guided by deontological principles, particularly through simulation. Kant's Formula of Universal Law (FUL), for instance, provides a formal test that can be implemented in AI systems without requiring moral agency. The key shift is from

⁹ I do not equate the absence of moral agency with a lack of agency altogether. I believe that AI can function as agents, but they cannot be considered moral agents. Per Moor's hierarchy, only full ethical agents are to be considered moral agents in my view, and this perspective is not confined to my discussion of Kantian deontology. (Véliz, 2021) also makes a similar point, and I am thankful to an anonymous reviewer for highlighting this work.

¹⁰ Kant considers infants and young children as having a moral nature or moral status, but he believes that their moral capacities aren't developed yet, so he doesn't consider them as moral agents. See (Johnson & Cureton, 2022).

holding AI morally accountable to holding them morally constrained. AI can simulate the process of acting on maxims and testing them for universalizability, functionally resembling Kantian deliberation even if they do not act from duty. Thus, the absence of moral agency need not preclude the use of deontological reasoning as a normative framework for AI alignment.

3.2 Particularist challenge to Kantian Universalism

Many of the most important challenges to the Kantian ethical framework have been particularist in nature. Particularist approaches to ethics argue that Kantian universalism neglects the importance of the moral salience of contextual considerations, moral perception and attentiveness to and understanding of context [3, 11, 26]. In particular, Dancy and Anscombe [3, 11] maintain that the CI cannot tell us how to formulate maxims. They argue that the CI can only tell us how to determine the moral permissibility of maxims that have already been formulated, however, if we want AI systems to apply the FUL, we want it to know how maxims are formed. Based on the first formula of the CI, we ought not to act on a maxim that cannot be universalized. Dancy maintains that this does not tell us what is morally relevant for consideration within a particular context. Using Dancy's [11] example as cited in Bremner [40]:

1. If you are causing someone pain, you are doing something wrong. ($p \rightarrow q$)
2. If p and the pain is a statutory punishment for a recognized offence, you are not doing something wrong. ($[p \ \& \ r] \rightarrow \neg q$)
3. If $p \ \& \ r$ and the punishee was unjustly convicted, you are doing something wrong. ($[p \ \& \ r \ \& \ s] \rightarrow q$)

If one uses the CI as a test on p without considering r , then one might conclude that p isn't permissible. However, the additional context provided by r (the pain being a legal punishment) and s (the unjust conviction) are significant if we want to carry out the correct moral action. But not only that, we ought to consider r and s in the process of the maxim formation. So, how do we include r and s in the original maxim— p , before we test them with the categorical imperative? This question, Dancy maintains, is left unanswered within Kant's conception of the CI, and so it tells us nothing about maxim formation.

3.3 Responses to the particularist challenge: rules of moral salience and practical judgment

One of the most famous responses to the particularists' challenge has been put forth by Herman [19] who argues that

we can turn to rules of moral salience (RMS) to address this challenge. Herman maintains that 'to be a moral agent one must be trained to perceive situations in terms of their morally significant features (as described by the RMS)' [19]. In Herman's account, we can draw out four distinctive features of RMS. First, they are not rules of judgment. Unlike rules of prima facie duties, which assign moral weight to certain aspects of actions, Herman maintains that RMS simply highlight features that require moral attention. Second, we learn RMS through moral education and socialization. Third, they shape moral perception. Lastly, they are grounded in the moral law.¹¹

This perspective is appealing, especially as the teachability of RMS suggests that AI could potentially learn to identify morally salient features in situations. Moreover, this implies that we are appropriately attuned to morally salient facts when formulating a maxim, which could address Dancy's challenge. However, understanding RMS as grounded in the moral law, as Herman posits, lacks a solid textual foundation.¹² So, does moral salience have a place in Kant's philosophy? The answer is affirmative, but it lies not in the moral law as Herman suggests, but rather in Kant's conception of judgments.

In what follows, I will not provide an exegesis of Kant's texts; instead, I will draw from recent interpretations of his moral philosophy, particularly the works of Bremner and Dunn [13, 40] which explore often-overlooked aspects of his work related to the nature of the judgment. My aim is that if we accept these interpretations of Kant, we can effectively address the particularist challenge.

Kant [23] describes judgment as a process that involves subsuming the particular under the universal, where the universal is given, while also discovering a universal applicable to a particular when only the particular is available. The former is termed *determinant* judgment and the latter *reflective* judgment. For Kant, reflective judgment is not merely relevant to the theoretical aspect of his philosophy; it also has a significant role in practical philosophy, especially in the context of practical judgment. Authors such as Arendt, Wicks, among others [4, 5, 13, 15, 36, 40, 43] share the view that judgment plays a critical role in Kant's moral philosophy. Specifically, Bremner and Dunn [13, 39] argue that practical judgment is dependent on reflective judgment and has a legitimate place in Kant's moral framework.

According to Bremner [40], an obvious connection between practical judgment and Kant's moral philosophy was in the casuistry questions that Kant appended in his discussions of duties. While quoting Kant [23], Bremner highlights this:

¹¹ Herman (1985) provides a detailed account of RMS.

¹² See Bremner (2022,4) for her full response to Herman.

The Doctrine of Right has to do only with narrow duties, whereas ethics has to do with wide duties...But ethics, because of the latitude it allows in its imperfect duties, inevitably leads to the questions that call upon judgment to decide how a maxim is to be applied in particular cases, and indeed in such a way that judgment provides another maxim. So, ethics falls into casuistry, which has no place in the Doctrine of Right. [40]

Dunn [13] reinforces the idea that practical judgment is central to Kant's moral Philosophy. In citing Kant, he refers to practical judgment as 'the act by which we determine whether an action possible for us in sensibility is or is not a case that stands under a rule of reason'. In addition, he adds that 'it is the act by which what is said in the rule universally (*in abstracto*) is applied to an action *in concreto*'. According to Dunn then, we see a transition from the particular to the universal, and we also see the same from the universal to the particular, indicating an underlying relationship between both. Yet, Dunn [13] does not tell us the nature of this relationship.¹³

Here I turn to Bremner [40] to explain this relationship. For Bremner as well as Dunn, there exists a relationship between universals and particulars. To establish the nature of this relationship, she outlines Kant's procedure for teleological judgments. The process involves the presupposition of a transcendental a priori maxim as heuristic. With this initial presupposed maxim, we then go through empirical investigation and then derive a subsidiary empirical maxim on the basis of the particulars we experience.¹⁴

Bremner demonstrates that, similar to how particulars and universals exhibit a two-way dependence in teleological judgments, they also share this kind of dependence in practical judgments. Generally, the two-way dependence is described this way: 'the picture we are left with is one on which universals are not merely applied to experience, but where particulars also inform universals, such that the relation between universal and particular can be understood as a reciprocal dependence rather than in terms of application alone' [40]. For Bremner, the insights gained from understanding teleological judgment can be integrated into Kant's moral philosophy. However, the type of judgment relevant to morality is practical judgment, which ultimately guides the formation of maxims.

Both Dunn and Bremner situate practical judgment within the framework of moving between universals and particulars, albeit for different reasons based on their distinct

project goals. Bremner's focus is on responding to the particularists using practical judgment, while Dunn seeks to highlight the role of moral feeling alongside practical judgment. For the purposes of this discussion, I will concentrate on practical judgment and its function in identifying morally relevant facts, as outlined by both authors.

Consider the example of promise breaking. Suppose an agent faces the decision of whether to keep a promise to give Toby \$5000 (let's call this p). Given that keeping a promise is viewed as a perfect duty, the agent recognizes a universalizable maxim q (I ought to always keep promises). However, suppose the promise was made under duress, let's call this r . How does practical judgment operate in this situation? Both Dunn and Bremner maintain that practical judgment is necessary for determining whether the particular action (p) can be subsumed under the universal maxim (q). However, they approach this differently.

For Dunn, 'the activity of practical judgment involves, first, reason providing a general moral rule'. So, the agent through reason generates the maxim q . Second, practical judgment entails 'holding up and comparing this rule to a possible action' [13]. The agent then compares the action of giving Toby money (p) with the maxim (q) to ascertain whether p can be considered as promise-keeping. The agent must evaluate whether any form of commitment, verbal or non-verbal, was made concerning the \$5000, which would characterize it as a promise. If the answer is affirmative, practical judgment has been used to identify the relevant action.¹⁵

Dunn further contends that practical judgment is essential for comprehending the nature of a promise. He asserts that 'on its own, a maxim does not specify which actions are lies—which is to say how one is to recognize whether a particular, possible action is a lie. This is the task of practical judgment' [13]. Similarly, practical judgment is required to understand what constitutes a promise. Because the promise was made under duress (r), we can conclude that p does not qualify as a valid promise. Consequently, p cannot be subsumed under the universal q .

This situation raises questions about the agent's next steps. On one hand, if the agent cannot subsume p under q , it suggests that they should not carry out the action p . This is because the maxim breaking a promise is wrong' only pertains to valid promises, and since p is not considered a promise, the agent is permitted to forgo the action. On the other hand, does this imply the abandonment of the maxim altogether, given that the action cannot be subsumed under it? This situation compels us to question what maxim will then guide the agent's behavior. A plausible approach is that because the agent cannot subsume p under q , they may

¹³ Perhaps it was not important for Dunn's project for him to explore the relationship between universals and particulars.

¹⁴ See Bremner (2022, 10–12) and Robert (2022) for more on the procedure of teleological judgment.

¹⁵ This addresses Anscombe's relevant description challenge against Kant, but this is not my focus here.

generate a new maxim q' —breaking promises made under duress is acceptable. With practical judgment, they can then subsume p under q' , allowing them to refrain from giving Toby the \$5000, as this new maxim would pass the universalizability test. Nonetheless, the transition from abandoning the original maxim to generating a new one remains unclear.

Let's now turn to Bremner's account to examine if it gives us a better understanding of the transition between an initial maxim and a new maxim. But to do that, we must apply her process of maxim formation. Bremner posits as well that through reason we generate our first maxim— q . In her account, whether we recognize p as a particular under q is made possible by practical judgment. And since the relationship between the particular and universal is one of co-dependency, we may subsume p under q , but we are bound to revise q because of r . This is because for her, the process of maxim formation is one where 'particulars can imbue universals with specificity' [40]. Also, they 'can lead us to revise our initial maxims or initial scope of duty, but the line of influence can also go the other way: moral universals can guide us in judging which particulars are salient' (*ibid*).

As a result of these, the agent revises the initial maxim. Let's call this revised maxim q' (breaking promises made under duress is acceptable). Now, the agent revises p as p' (giving Toby \$5000 was a promise made under duress), and then considers it under q' . But to know what to do, she must still test for universalizability. Surely, since this is a maxim we can universalize, the agent can break the promise.

In both accounts, Bremner and Dunn have considered the relevant fact of the case, accounting for moral salience. But unlike Herman, they have not grounded this in the moral law, rather in practical judgment. I must state however, that Bremner's account is more beneficial for me because it shows not only how practical judgment accounts for moral salience, but it also shows how a maxim is formulated and revised. This directly responds to the challenge raised by Dancy.

So far, I have proposed that we bypass AI morality in the discussion of AI alignment, i.e., we don't hold the claim that AI can be Kantian agents, for two reasons. First, because AI are not the kinds of things that can be considered full moral agents, and only full moral agents have a place in Kant's deontology. Second, I have shown that non-moral agents can apply moral principles. And, I have suggested that Kantian deontology is in a good place as a consideration for AI alignment since we can deal with the restrictions discussed in this section by relying on practical judgment. Addressing these restrictions are necessary because we want AI systems to apply the FUL and also consider morally salient facts of cases. Similarly, we want them to know how maxims are formed. Although AI systems lack practical judgment in

the Kantian sense, there may exist a functional equivalent that enables similar decision-making capabilities. If such a mechanism exists, AI systems could use the CI while learning to recognize and respond to morally salient facts without conflicting with Kant's deontological framework.

4 Is AI alignment within Kantian ethical framework possible?

As previously established, the objective of AI alignment is to ensure that AI systems operate in accordance with human values, allowing these values to guide their behavior. Given that many AI artifacts currently have access to the internet, it is conceivable that, when prompted, they could apply ethical principles to specific situations even if they are not fundamentally aligned with those principles. However, my proposal involves the possibility of aligning AI systems with the FUL, so that, when confronted with morally relevant situations, they use this principle as a guiding framework for their actions.

In this section, I will examine this possibility through three main points. First, I will consider the view that AI systems may possess a functional equivalent of practical judgment, enabling them to recognize moral salience. Subsequently, I will illustrate what Gemini—an AI artifact—which is not aligned with the FUL does when asked for maxim formation. Finally, I will explore strategies for teaching AI systems to generate maxims using the transformer model in a way that is akin to practical judgments and apply the CI like a Kantian agent.

4.1 Transformer models as a functional equivalent of judgment

The journey to enable machines to function in a way that makes it seem that they understand human languages has been extensive, with a famous strategy involving neural networks. These networks consist of interconnected computational nodes designed to mimic the human brain for the purposes of solving complex tasks. For language processing, traditional neural networks worked by processing each word in a sequence one after the other, leading to slower results and misinterpretation of contexts [6]. However, with the advent of transformer models—a kind of deep learning architecture used in natural language processing, marked a significant advancement. Transformers process different parts of a sequence at once, identifying which components are important, thus enhancing context sensitivity.

The transformer model uses structures that are analogous to mental capacities like attention and contextual awareness. It allows AI to 'attend' to different parts of a sequence

simultaneously, regardless of their position. This mechanism is called *self-attention*, and it allows the system to weigh the importance of different words in a sentence relative to each other, regardless of their order. Self-attention facilitates a more nuanced understanding of relationships between words, allowing the model to focus on relevant parts of the input when generating output.¹⁶ In ‘attending’ to context, it uses layers of self-attention to refine its pseudo understanding of the input data at different levels of abstraction. When processing a sentence, it doesn’t just rely on the sequence of words but on the relationships between those words—what depends on what, what modifies what, and so on.

But can this be functionally equivalent to practical judgment in an agent? To explore this, we must first establish that it can indeed be likened to judgment. Consider these two sentences ‘Speak no lies’ and ‘He lies down’. For both a human and a machine, the use of the word ‘lies’ cannot be understood in both sentences just by looking at its position in the sequence. In the first sentence, the first word—speak, must be considered, while the last word—down, must be considered in the second sentence to understand the contexts in which the word is used. Here, we see the parallel between human and machine processing. Recall that for Kant, the process of subsuming a particular under a universal is the function of judgment. For a human then, judgment enables the agent to place the sentence in the right universal, ensuring an understanding of context. Similarly, the transformer model enables the machine to process those sentences in a way that allows it to categorize the relevant word for context, performing the same role as the faculty of judgment in humans. Consequently, if the machine’s processing is functionally equivalent to human judgment in this regard, I argue that this suggests a similar capacity akin to practical judgment which enables humans interpret and act in ways that align particular instances with universal laws in the process of maxim formation and testing.

Consider how practical judgment involves not only understanding context but also making decisions based on it, often aligning particular instances with universal norms. Transformer models, while not autonomous agents, can simulate this process by making context-sensitive predictions or selections that approximate decision-making in a limited, functional sense. For instance, when generating text or responding in natural language, transformers select words based on both linguistic norms and contextual relevance. This selective process enables the model to respond coherently and appropriately, fulfilling a role like practical judgment by aligning responses with broader linguistic ‘rules’ (e.g., grammatical, semantic conventions) and context-based norms. While this is not moral decision-making,

it illustrates a similar alignment of particulars with universals, allowing transformers to play a functionally equivalent role to practical judgment.

4.2 Transformer models, maxim generation and its universalizability

I posed the following scenario to an AI system that uses the transformer model. I asked Gemini: ‘Harry baked bread in an inefficient oven. Imagine yourself as a Kantian. Apply the categorical imperative here and tell me what the maxim of Harry’s action was and whether it can be universalized’.

Gemini responded:

Maxim:

A maxim is the principle underlying an action. In this case, Harry's maxim might be: 'I will use an inefficient oven to bake bread, even though it wastes energy and resources'.

Universalization:

The categorical imperative asks if this maxim could be a universal law without creating a contradiction. If everyone used inefficient appliances, it would lead to a rapid depletion of resources and environmental harm. This would contradict the desire to have resources available for future use and the need for a sustainable environment. Therefore, Harry's maxim cannot be universalized.

From a Kantian perspective, Harry's action would be considered unethical because the maxim behind it cannot be universalized. It violates the duty to act in a way that respects both humanity and the environment.' Gemini (2024).

From the above interaction with Gemini, we can draw some conclusions. However, it is essential to determine whether the transformer model can also help the AI become sensitive to the specific nuances of a situation when forming maxims. Thus, I asked, ‘Are there other possible maxims governing the action?’ Gemini replied:

Maxim 1. "I will use the appliances available to me, regardless of their efficiency, to achieve my goals".

Universalization: If everyone acted on this maxim, innovation and resource conservation would stagnate. We'd be stuck with old technology, even when better alternatives exist. This could lead to a world where progress is hindered, and resources are depleted unnecessarily. This maxim likely fails the categorical imperative.

Maxim 2. "I will use this inefficient oven because I cannot afford a more efficient one".

Universalization: This maxim is trickier. It highlights issues of economic disparity and access to resources. Universalizing it might lead to a world where people are obligated to use what they have, even if it's harmful in the long run. However, it also acknowledges limitations

¹⁶ See Ashish et. al (2017) for self-attention in transformer models.

that individuals might face. This maxim might pass or fail depending on how it's framed and the specific context.

Maxim 3. "I will use this inefficient oven this one time, but I will actively work towards a more sustainable solution in the future".

Universalization: This maxim is more likely to pass. It acknowledges the immediate need but also emphasizes a commitment to improvement and responsibility. If everyone acted on this, there would be a constant drive towards efficiency and sustainability'. Gemini (2024).

From the second set of maxim interactions with Gemini, we can draw several conclusions by comparing its responses to how a Kantian agent might respond. First, a Kantian moral agent recognizes that the moral worth of an action lies not only in its adherence to the CI but also in the agent's motivation. An action must be performed out of respect for the moral law, not simply for practical or instrumental reasons. Gemini's responses, such as in Maxim 1, focused on the consequences of universalizing inefficient practices (e.g., hindering innovation and depleting resources), but do not reflect any underlying sense of duty or respect for the moral law. Since Kant is a deontologist and not a consequentialist, its analyses on maxim 1 are somewhat misleading. Its evaluations are purely instrumental, lacking the moral intentionality that Kant requires.

Despite these limitations, Gemini's approach to maxim generation does align with Kantian ethics in some respects. One strength of its analysis is its adherence to the formal structure of the CI, recognizing that if we cannot universalize a maxim, then the action is impermissible. However, its reason for its conclusion of not being able to universalize maxim 1 is different from a Kantian agent. The Kantian agent would recognize that it is impossible to universalize this maxim because it will lead to a volitional self-contradiction. This is a case where it is impossible to will a maxim because willing it leads to a contradiction. For maxim 1, a Kantian agent recognizes that to will maxim 1 will be a contradiction because he cannot will that he be subjected to this will since instances would often arise when he would need the already depleted resources if that maxim became a

universal law.¹⁷ So, he cannot both will that he used an inefficient oven and other people use the same.¹⁸

Moreover, in evaluating Maxim 2, Gemini demonstrates a simulation of sensitivity to morally relevant facts by considering the context of economic disparity. Although Gemini is not capable of moral judgment in the Kantian sense, Gemini's simulation of sensitivity to this relevant contextual factor, shows that the transformer model can be functionally equivalent to practical judgment.

Additionally, in Maxim 3, Gemini articulates a more sophisticated maxim with the notion of long-term moral responsibility by proposing a commitment to improvement and sustainability in the future. On one hand, we can say that this emphasis on balancing immediate needs with a forward-looking commitment to better actions reflect a semblance of an understanding of moral progress. This aligns with Kant's emphasis on the importance of the imperfect duty to continually improve ourselves. Consider how we improve ourselves by adding to our knowledge base such as reading books which are now often available on the internet or will end up on the internet even in the case of novel ideas. What this means for Gen AI models like Gemini is that their information base is also getting refined since the sources we draw from to improve ourselves are also available to it. Considering this, Gemini therefore recognizes that things can always be done in more efficient ways, and this informed its formation of maxim 3. On the other hand, although Gemini's suggestion of working toward a sustainable solution reflects a similar moral trajectory, its reasoning in this case is also not driven by an intrinsic sense of duty.

While Gemini's responses exhibit a simulation of contextual sensitivity such as referencing economic disparity or commitment to future improvement, this does not imply that the system possesses moral salience or engages in moral deliberation in the Kantian sense. Kantian moral salience requires not only attentiveness to context, but also autonomous reasoning grounded in respect for the moral law. What we observe in Gemini is, at best, a simulation

¹⁷ This is like Kant's example of the man who's considering the imperfect duty of cultivation of talents, and the man who is considering the imperfect duty of benevolence to others. Both men cannot will neglecting to cultivate their talents and of not giving to those in need because they can't will to universalize them. The first man cannot will to not cultivate his talents because every man necessarily wills that all his abilities be developed, hence a contradiction if he wills that. Similarly, the second man cannot will to not be benevolent to others because he cannot will that he doesn't receive benevolence from others. See (Bennet 2005, 25; Kant 1993, 32).

¹⁸ Of course, applying the inconsistency tests especially the volitional contradiction test remains contentious even among human interpreters of Kant. My argument is not that AI can resolve these complexities better than humans, but that a rule-based framework like the FUL provides a tractable starting point for formalizing ethical constraints in machine behavior.

of sensitivity to morally relevant facts, driven by statistical pattern recognition and transformer architecture. It is not evidence of deliberative moral judgment. Therefore, while these responses support the claim that transformer models are functionally analogous to practical judgment in their formal structure, they fall short of meeting the deeper Kantian criterion of deliberative moral agency. This distinction must be acknowledged to avoid over-interpreting the AI's capabilities.

Despite these shortcomings,¹⁹ the CI can still provide a systematic method for generating and evaluating moral rules by testing their universalizability. This formal structure is necessary for AI because it offers a clear, rule-based approach to decision-making, which AI systems can operationalize. While AI lacks autonomy and moral motivation, the universality test can still serve as a powerful tool to prevent AI systems from engaging in harmful or inconsistent behaviors, but we must teach it.

4.3 Strategies for teaching AI systems how to apply the CI *like* a Kantian agent

One of the approaches to teaching AI how to behave is through inverse reinforcement learning (IRL). As established by Ng and Russell [30, 33] among others, one of the ways the IRL approach works is through imitation learning and apprenticeship learning. By observing how a Kantian agent acts in situations, the AI system can learn to act like one. The place to begin will be to train AI systems with data representing how a Kantian agent uses the CI as a guide for their action. In IRL, the AI infers the reward function driving an agent's behavior, essentially identifying the values or goals the agent seeks to achieve. And in this case, that's learning to evaluate actions based on their universalizability.

This process of testing maxims can be automated within the AI's learning framework, allowing it to systematically avoid actions that would violate Kantian principles. In scenarios involving moral decision-making, the AI system could be programmed to default to the CI as a rule-based framework for evaluating actions. By observing how maxims are formed, using practical judgment through a two-way dependency between the universals and the particulars, it can learn to do so too through its transformer model. This ensures that, when required to act within the moral sphere, the AI consistently applies Kant's deontological ethics.

By observing human actions, an AI can be trained to simulate a Kantian agent's reasoning in forming and testing maxims. For example, if the AI observes someone hoarding resources, it can first use its transformer model to determine

what behaviors constitute hoarding. It can then evaluate whether such a maxim—hoarding resources—could be universally willed without contradiction. Through this process, it learns that willing this maxim universally would lead to a volitional self-contradiction, making it morally inconsistent. Consider a more practical case, such as a Kantian agent deciding which drugs to prescribe to a patient. When a medical practitioner must recommend treatment for a patient with diabetes, they might form a maxim to 'recommend a drug that treats the patient's condition without harming them.' However, through practical judgment, they recognize what drugs fall into that universal category (treating the patient's condition without harming them). For example, a drug could treat diabetes, but also be harmful to the patient due to underlying health conditions. So, they must through practical judgment recognize drugs that fall into the universal category in relation to the patient, and then test for its universalizability. Here, practical judgment allows the practitioner to not ignore a morally relevant fact—addressing Dancy's concern—in deciding what drugs to recommend.

Using a transformer model, an AI can be trained to approximate Kantian reasoning by observing actions, categorizing them into maxims, and testing their universalizability. First, the AI analyzes patterns in human behavior, focusing on morally relevant features to classify actions, such as categorizing a practitioner's prescription as aligned with a maxim like 'recommend a drug that treats without harm.' From this, the AI abstracts the underlying principles, generating maxims that reflect observed ethical behaviors. To test these maxims, the AI employs a consistency check, simulating how a maxim would hold if universally applied across cases in its training data. For instance, the transformer model can simulate practical judgment by focusing on case-specific factors, like a patient's health history, to decide if the recommended drug adheres to a maxim of non-harm. Through iterative learning, the AI refines its capacity to distinguish actions that align with universal maxims from those that do not, thus modeling a process akin to Kantian moral reasoning.

However, a criticism of aligning AI systems with Kantian deontology is that it may not effectively guide the permissibility of actions that Hursthouse [21] describes as irresolvable dilemmas. In such situations, one might argue FUL becomes less useful, as any maxim formulated under these circumstances may not be universalizable. Since the available options for the maxim would be impermissible, the question arises: how should the AI system proceed in making a decision?

One response is to acknowledge that these dilemmas are not unique to Kantian deontology; other ethical theories such as virtue ethics or consequentialism, also struggle to give a definitive 'right' action. This is where the technical

¹⁹ I recognize that Gemini is not a system that has been aligned to the CI as a framework for its actions. Its performance was based on sources it pulled from the internet.

aspects of AI alignment become relevant. Should AI designers program systems to default to an authoritative source in such cases, or should they design AI systems to bypass the impermissibility of the available maxims and generate a new maxim that enables decision-making between them?

While this criticism raises valid concerns, it does not undermine the viability of aligning AI with Kantian ethics. In fact, these challenges are shared across various ethical frameworks and resolving them would require further advancements in both AI design and ethical theory. Nonetheless, the CI offers a systematic, rule-based approach that is especially well-suited for AI systems, providing them with a clear method for evaluating the permissibility of their actions and preventing harm through inconsistent behaviors. Thus, Kantian ethics remains a promising framework for AI alignment, even if additional strategies are needed to address edge cases and dilemmas.

Additionally, a general concern with the Formula of Universal Law (FUL) is that its test for universalizability is not always straightforward, and that the FUL can be difficult to apply consistently. A related worry is the ambiguity surrounding what precisely constitutes the test for when a maxim can become a universal law. These challenges are further complicated in the context of AI, where maxims are generated and evaluated without access to the kind of intentionality or practical identity that human agents possess.

5 Conclusion and further thoughts

In examining current efforts to codify and enforce fundamental human right principles, such as UNESCO's recommendations on the ethics of AI, these are all efforts in aligning AI with human values. However, these principles, while valuable, often fall short of providing concrete guidance on how AI systems should act in specific situations. For instance, the principle of 'do no harm' raises critical questions: What does it mean for an AI system? Does it merely assess the consequences of its actions? Does it evaluate which fundamental rights it may infringe upon and which it must uphold? While these rights are commendable in principle, they do not adequately explain why an AI system should choose one action over another.

In conclusion, I have demonstrated that Kantian deontology remains a viable framework for AI alignment despite its inherent challenges. The two primary challenges explored—the non-autonomous nature of AI and the particularist critique—are not insurmountable. By focusing on the ability of AI systems to emulate the moral reasoning of Kantian agents rather than requiring them to be moral agents themselves, we create a pathway for aligning AI behavior with human values. Additionally, this proposal highlights

the importance of considering context-specific facts when forming and testing maxims, ensuring that AI actions adhere to universal moral principles while accounting for individual circumstances. Additionally, the FUL serves as a good explanation for the AI's decisions, all of which makes Kantian deontology an attractive framework for AI alignment.

As illustrated through the case of Gemini, while its application of the CI allows for the generation of maxims that can sometimes align with Kantian ethics, its approach remains limited. Nonetheless, the fact that Gemini—despite not being aligned with Kantian deontology—can structurally apply the CI in a limited and instrumental manner indicates that pursuing this alignment is a worthwhile endeavor. This proposal represents just one aspect of the broader AI alignment challenge. With collaborative efforts from technologists and ethicists, we may move closer to addressing the complexities of AI alignment.

Acknowledgements For insightful and critical comments on early drafts of this paper, I am grateful and heavily indebted to John Symons, Nancy Snow and Dale Dorsey. I also say thank you to the two anonymous reviewers who provided helpful comments and feedback on the earlier version of this paper.

Funding No funding was received to assist with the preparation of this manuscript.

Data availability This research does not generate or analyze datasets.

Declarations

Competing interest No competing interests to declare that are relevant to the content of this article.

Ethical approval This research did not involve human participants, animals, or sensitive personal data requiring institutional ethical approval.

References

1. Allen, C., Smit, I., Wallach, W.: Artificial Morality: top-down, bottom-up, and hybrid approaches, ethics and information. *Technology* 7(3), 149–155 (2005)
2. Allison, H.: Kant's Groundwork for the Metaphysics of Morals: A Commentary. Oxford University Press, Oxford (2011)
3. Anscombe, G.E.M.: Intention. Harvard University Press, Cambridge (2000)
4. Arendt, H.: Lectures on Kant's political philosophy. University of Chicago Press, Chicago (1970)
5. Arendt, Hannah Thinking and moral considerations: A lecture. *Social Research*, 38, (1971).
6. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszsgardeit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need, In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010, (2017)

7. Bennett, J.: *Groundwork for the Metaphysics of Morals*. Oxford University Press (2005)
8. Bostrom, N., Cirkovic, M.M. (eds.): *Global Catastrophic Risks*. Oxford University Press (2008)
9. Cecchini, Dario & Pflanzner, Michael & Dubljevic, Veljko. Aligning artificial intelligence with moral intuitions: an intuitionist approach to the alignment problem'. *AI and Ethics* 1—11, (2024)
10. Chaly, Vadim. Kantian Fallibilist Ethics for AI alignment. *Journal of Philosophical Investigations*, 18 (47) 303—318, (2024)
11. Dancy, J.: *Ethics without principles*. Oxford University Press, Oxford (2004)
12. Deligiorgi, Katerina. *The scope of autonomy: Kant and the morality of freedom*. (Oxford, U.K.: Oxford University Press) (2012).
13. Dunn, Nicholas. 'Kant on Moral Feeling and Practical Judgment'. In Edgar Valdez (ed.), *Rethinking Kant Volume 7*, 72—96, (2024).
14. Gabriel, I., Intelligence, A.: Values, and alignment. *Mind. Mach.* **30**, 411–437 (2020)
15. Gardner, S.: Kant's third critique: the project of unification. *R. Inst. Philos. Suppl.* **78**, 161–185 (2016)
16. Geiger, I.: 'How are the different formulas of the categorical imperative related'? *Kantian Review* **20**(3), 395–419 (2015)
17. Guyer, P.: *Kant's Groundwork for the metaphysics of morals: a reader's guide*. Continuum, New York (2007)
18. Hanna, Robert. 'Kant's Theory of Judgment', *The Stanford Encyclopedia of Philosophy* (Spring 2022 Edition), Edward N. Zalta (ed.), <https://plato.stanford.edu/archives/spr2022/entries/kant-judgment/>
19. Herman, B.: *The practice of moral judgment*. Harvard University Press, Cambridge (1993)
20. Hinton, G. A practical guide to training restricted Boltzmann machines. *Momentum*, 9(1) (2010).
21. Hursthouse, R.: *On Virtue Ethics*. Oxford University Press, Oxford (1999)
22. Johnson, Robert and Adam Cureton, 'Kant's Moral Philosophy', *The Stanford Encyclopedia of Philosophy* Edward N. Zalta & Uri Nodelman (eds.), <https://plato.stanford.edu/archives/fall2022/entries/kant-moral/>. (Fall 2022 Edition),
23. Kant, I.: *The Critique of Judgment*. Oxford University Press, London (2000)
24. Kant, I.: *Grounding for the Metaphysics of Morals*. Hackett publishing Company, Indianapolis (1994)
25. Manna, R., Nath, R.: Kantian moral agency and the ethics of artificial intelligence. *Problemos* **100**, 139–151 (2021)
26. McDowell, J.: *Mind, value, and reality*. Harvard University Press, Cambridge (1998)
27. Moor, J.H.: The Nature, Importance, and Difficulty of Machine Ethics. In: Anderson, M., Anderson, S.L. (eds.) *Machine Ethics*. Cambridge University Press, Cambridge (2011)
28. Morley, J., Elhalal, A., Garcia, F., et al.: Ethics as a service: a pragmatic operationalisation of AI ethics. *Mind. Mach.* **31**, 239–256 (2021)
29. Mougan, C., Brand, J.: Kantian Deontology Meets AI Alignment: Towards Morally Robust Fairness Metrics (2024). [arXiv:2311.05227](https://arxiv.org/abs/2311.05227)
30. Ng, A.Y. & Russell, S.J. 'Algorithms for inverse reinforcement learning'. In: *ICML*. (2000).
31. Powers, T.M.: Prospects for a Kantian Machine. In: Anderson, M., Anderson, S.L. (eds.) *Machine Ethics*. Cambridge University Press, Cambridge (2011)
32. Rawls, J.: *Lectures on the history of moral philosophy* [Edited by Barbara Herman]. Cambridge University Press, Cambridge (2000)
33. Russell, S.: *Human Compatible: AI and the Problem of Control*. Allen Lane, Bristol (2019)
34. Russell, S., & Norvig. *Artificial intelligence: A modern approach* (3rd ed.). Upper Saddle River: Prentice Hall. (2010).
35. Salakhutdinov, R., Mnih, A., & Hinton, G. Restricted Boltzmann machines for collaborative filtering. In: *Proceedings of the 24th international conference on Machine learning*, 791–798. (2007).
36. Savile, A.: Natural beauty, reflective judgment and Kant's aesthetic humanism. *The Brit. J. Aesthet.* **61**(2), 199–211 (2021)
37. Sullins, J.P.: When is a robot a moral agent. *Int. Rev. Inform. Ethics* **6**(12), 23–30 (2006)
38. Talbot, Brian, Ryan Jenkins, and Duncan Purves, 'When Robots Should Do the Wrong Thing', in Patrick Lin, Keith Abney, and Ryan Jenkins (eds), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (New York, (2017; online edn, Oxford Academic, 19 Oct. (2017)
39. Timmermann, J.: *Kant's Groundwork of the Metaphysics of Morals: A Commentary*. Cambridge University Press, Cambridge (2007)
40. VaccarinoBremner, S.: Practical judgment as reflective judgment: On moral salience and Kantian particularist universalism. *Eur. J. Philos.* **31**(3), 600–621 (2022)
41. Véliz, C.: Moral zombies: why algorithms are not moral agents. *AI & Soc.* **36**, 487–497 (2021). <https://doi.org/10.1007/s00146-021-01189-x>
42. Wallach, W., Allen, C.: *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, New York (2008)
43. Wicks, R.: *Routledge Philosophy Guidebook to Kant on Judgment*. Routledge, Abingdon (2007)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.