Journal of Politics and Ethics in New Technologies and AI

Vol 4, No 1 (2025)

Journal of Politics and Ethics in New Technologies and AI

Distributed Virtue and the Phantom Agent: Rethinking Moral Responsibility in Human–AI Systems

Serap Keles

doi: 10.12681/jpentai.41895

Journal of Politics and Ethics in New Technologies and AI
Volume 4, Issue 1 (2025)
e-ISSN: 2944-9243

https://doi.org/10.12681/jpentai.41895

RESEARCH ARTICLE

Distributed Virtue and the Phantom Agent: Rethinking Moral Responsibility in Human–AI Systems

Serap Keles
Equity and Inclusion Lead, Bird College, Sidcup, UK.

Abstract
Advances in artificial intelligence (AI) challenge traditional notions of moral agency and responsibility. This paper introduces the concept of the phantom agent as an ontological and ethical category for AI systems that are neither mere tools nor full moral agents yet decisively shape moral outcomes in human–AI collaborations. Drawing on classical moral philosophy and engaging contemporary philosophy of technology, the analysis reframes responsibility in socio-technical systems. It argues for *distributed virtue*, an extension of virtue ethics to human–AI collectives, and examines *epistemic asymmetry*, specifically the uneven distribution of knowledge and transparency between human and AI as a central moral challenge. The paper defends an original account in which moral responsibility is reconceived as an emergent and shared property that human–AI systems exhibit traits of character and accountability distributed across their components. This approach aims to integrate ethical influence of AI systems (the phantom agents) into a coherent model of responsibility and virtue by moving the debate beyond existing paradigms.

Keywords: Ethics of AI, Distributed Virtue, Phantom Agent, Moral Responsibility, Ethics of Human–AI Systems

Introduction

The growing ubiquity of AI in everyday life manifest in autonomous vehicles navigating public roads, algorithms shaping access to social welfare and digital assistants influencing clinical diagnoses has brought to the fore a question both urgent and unresolved: Who, or what, bears moral responsibility for the actions and consequences produced by human–AI systems? This question resists easy answers, partly because it unsettles the conceptual architecture of our moral reasoning. Traditional ethical frameworks, rooted in anthropocentric assumptions, can falter when faced with entities that act, decide and adapt yet lack consciousness, intention and moral self-awareness.

One prevailing view casts AI systems as nothing more than tools. Sophisticated, yes, but ethically inert, akin to knives, hammers or thermometers, as devices whose consequences are fully attributable to their human creators and users. In this framework, all moral responsibility remains with the human hand, whether behind the algorithm's design or at the interface of its use. In contrast, a growing body

of thought struggles that as AI systems become more autonomous, adaptive and inscrutable they begin to resemble agents whose operations are sufficiently complex and consequential to warrant at least partial ethical consideration. Some even raise the unsettling possibility that, under certain conditions, such systems may merit a moral status approximating personhood (Gunkel, 2012, pp. 99–100).

Yet this binary (AI as mere instrument vs. AI as moral agent) has led us into a conceptual dead end. It frames the ethical ground in oppositions that obscure more than they clarify, forcing us to choose between implausible extremes. What this paper proposes is a reframing and invitation to see AI outside the lens of rigid categories, something as occupying a morally significant, if ontologically ambiguous, space. To that end, I introduce the concept of the "Phantom Agent", a term that captures the uncanny, quasi-agentive (Kasenberg et al., 2018; Semler, 2022) role that AI systems now play. These are entities that haunt the moral view, influence outcomes, mediate choices and co-author consequences, but lack the interiority, deliberative capacity and answerability that would render them full moral subjects.

The phantom agent, then, is not an ethical subject in its own right, but neither is it ethically neutral. It is a figure whose presence complicates the attribution of responsibility, precisely because it operates in the space between action and intention, between cause and judgment. To recognise the phantom agent is to acknowledge that our moral frameworks must stretch to meet the complexity of our technological entanglements and in this troubled ground agency is partial, relational and distributed that the ethics of the present must find its footing.

A further and pressing concern in human–AI interaction is what might be called epistemic asymmetry (Floridi & Taddeo 2016; Sparrow, 2016; Matthias, 2004; Gunkel, 2012; Zerilli et al., 2019), that is the structural disparity between human and artificial agents in their access to information, their ability to interpret it and their capacity to make that process intelligible. In contemporary AI systems, especially those powered by machine learning, the mechanisms of decision-making often unfold within layers of statistical abstraction that is inaccessible even to those who have built and deployed them (for most recent discussions on "black box" paradigm, see, Zhao et al., 2024; Hassija et al., 2024). These systems detect patterns at a scale and complexity that outpaces human cognition, yet they provide few, if any, windows into how particular judgments are made. The human operator navigating this area may be left with an output devoid of rationale and a decision offered without explanation. And while the system hums with computational precision, the human is asked to respond to trust, act and be responsible often without the tools to meaningfully evaluate what has been presented.

This condition gives rise to what might be called a "moral vacuum", where ethical accountability separates across multiple actors, such as developers, users and institutions, but adheres to none. When

errors occur, each party may signal elsewhere. The user did not fully grasp the algorithm's process, the designer assumed responsible use, the AI, lacking consciousness, remains beyond the reach of blame. What emerges is a responsibility gap (Demirtaş, 2025; Kiener, 2024; Königs, 2022; Methias, 2024; Santoni de Sio & Mecacci, 2021; Vallor &Vierkant, 2024), and possibly confusion and silence. And in this silence democratic values risk erosion. For without transparency, intelligibility and structures that allow humans to remain morally present within technological decisions, responsibility weakens. The task, then, is more than a technical one, it is moral and institutional to redesign systems so that those who act within them are not mystified and absolved. At its core, a just society does not demand perfection, but it does require those who make choices (human or hybrid) and can be accountable to moral values and collective responsibility.

Aristotle, Kant, Mill and Beyond

Introducing AI into moral contexts challenges traditional ethics profoundly. From a moral standpoint, I develop this paper on the idea of distributed virtue as an extension of virtue ethics beyond the individual, proposing that in systems where humans and AI interact closely, moral virtues (and vices) can become properties of the human–AI collective rather than of either component alone. Virtue ethics, since Aristotle, has focused on the character and habits of individual moral agents, seeing virtue as a disposition developed through habituation and practical wisdom. Here we explore how virtuous (or unvirtuous) behaviour might be cultivated or hindered by AI systems.

From a Kantian perspective, AI systems that lacking rational will and moral autonomy, fall outside the circle of moral agents. Kantian ethics urgently reminds us that humans bear the duty to be autonomous in decision-making processes. When humans defer blindly to algorithms, as when a judge uncritically adopts a sentencing recommendation, they risk surrendering their moral judgment, becoming instruments themselves and ruled subtly by computational logic rather than reason. In similar examples we see that Kant's emphasis on imputability further complicates responsibility in AI-mediated situations. A self-driving car's accident lacks a clear human will where programmers and passengers are not explicitly authoring the precise moment of harm, creating a responsibility gap. Kantian theory thus calls for heightened transparency and rigorous oversight, aligning with Coeckelbergh's (2020) relational concept of responsibility as answerability, where the human duty is to articulate clear reasons to those affected by algorithmic outcomes.

Mill's utilitarian framework (1969), by contrast, evaluates moral responsibility primarily through consequences, judging actions right as they enhance collective happiness. In algorithmic contexts, such as self-driving cars programmed to minimise total harm, utilitarian reasoning is compelling, directly

calculable and intuitively suited to technological logic. Yet consequentialism alone cannot fully satisfy moral intuitions that demand individual accountability, especially when isolated harms occur despite overall utility gains (Schmidt, 2024). Seen in this light, although utilitarianism guides practical rulemaking, such as mandating human oversight in critical AI decisions, it struggles with assigning moral responsibility beyond statistical reasoning.

Recognising these limitations, contemporary thinkers introduce new philosophical resources. Floridi's (2013) concept of distributed morality suggests moral responsibility as an emergent property shared across human-AI networks. Rather than isolating accountability within a single agent, Floridi proposes viewing morality as distributed among designers, deployers and technological artifacts themselves, each holding partial causal roles in outcomes. This approach aligns practically with modern regulatory proposals, such as algorithmic impact assessments that acknowledge multi-agent responsibility without diluting individual accountability. Further, Vallor (2016) enriches this perspective by advocating technomoral virtues, emphasising moral habits essential for human flourishing in technological contexts virtues like humility, honesty and courage adapted specifically to our digital age. Vallor underscores the necessity of cultivating epistemic humility and proactive engagement, counteracting technosocial opacity by equipping individuals and institutions with virtues suited to our collective life alongside increasingly sophisticated AI. Coeckelbergh (2012; 2020) complements this with relational ethics, arguing that moral responsibility emerges dynamically within relationships, practices and contexts rather than residing statically within isolated entities. Coeckelbergh's call for responsibility-as-answerability insists on transparent justifications to affected parties, highlighting continual moral vigilance and dialogue throughout an AI system's lifecycle.

Bringing these contemporary perspectives together we see two common themes arises here:

- agency is not black-and-white,
- morality can be systemic and shared.

These insights inform the next part of the paper, where I build the case for phantom agency and distributed virtue explicitly. The phantom agent is proposed as a useful conceptual handle for the type of quasi-agent that AI represents. Acknowledging it helps us to clarify how virtue and responsibility can be allocated in a human–AI system. Further illustrating by using recent examples we will have concrete grounds to test these ideas.

Phantom Agents in Human – AI Collaboration

A phantom agent refers to an AI system (or generally, a technological system with AI components) that participates in achieving outcomes within a human–AI partnership in a way that mimics agency,

without fulfilling the criteria of full moral agency. It is phantom-like because it presents an appearance of independent decision-making and goal-directed behaviour (the hallmarks of an agent), yet it lacks a tangible moral self. These systems act in ways that shape outcomes, alter trajectories and influence human judgment, but they remain incorporeal within our ethical vocabulary. They are neither fully absent nor fully present in the moral field yet occupy a space of consequential ambiguity. Naming them as phantom agents enables us to acknowledge that something has indeed occurred (an act with ethical weight) without conflating that act with the agency of a moral subject in the human sense.

Phantom agents do not arise spontaneously. They are authored, designed, trained and deployed through layers of human decision-making. Their functional goals, even when they evolve in complex environments, remain attached to human-defined objectives. To that end, a chess engine programmed to win will pursue that goal with remarkable strategic sophistication and may even generate solutions unforeseen by its designers. Yet the telos (the end it pursues) has been conferred by human intention. This gives rise to a mode of agency that operates with operational independence but lacks interpretive self-awareness. AI systems exhibit what might be described as functional autonomy, which is the capacity to execute tasks, respond to stimuli and modulate behaviour in real time without direct human prompting. This capacity is neither trivial nor purely mechanical, however, through machine learning techniques these systems interact dynamically with their environments, refining outputs in ways that exceed the predictability of traditional tools. We can still say that despite this sophistication, such systems do not constitute moral agents in the Kantian sense as they do not reason through norms or deliberate about ends. As Floridi and Sanders (2004) suggest, their responsiveness, interactivity and adaptiveness mark a significant threshold in technological agency without moral autonomy. So that these systems can act with purpose-shaped trajectories without the internal structure of moral reflection.

This ambiguity perhaps helps us here to explain why humans are so prone to anthropomorphise AI. When a system behaves with consistency toward a goal, it begins to resemble something animate and intentional. It becomes easy for us to imagine the presence of a will, even where there is none. The phantom agent emerges here, not simply as a metaphor, but as a conceptual necessity for grappling with a form of agency that is entangled with human aims, operates with partial independence and yet remains ethically elusive. This in a way means that there is no conscious awareness, no capacity for guilt or empathy, no understanding of the moral meaning of its actions. It cannot reflect on the maxim of its action like a Kantian agent or cultivate virtue as an Aristotelian. It is, in a phrase, mindless agency (Floridi & Sanders, 2004). In this sense AI is a phantom agent that has not granting with personhood

but acknowledging the as-if agency in its behaviour while maintaining that morally it is more akin to an instrument.

Despite lacking moral consciousness, they can represent a real causal power in morally important events. For instance, an autonomous vehicle's decision algorithms causally determine whether the car swerves or brakes, so life and death could hinge on that. In the same vein, a content recommendation algorithm on a social media platform causally influences what information people see, potentially affecting opinions, elections, mental health and clearly morally charged various outcomes. In these scenarios, the AI's "choices" can lead to consequences typically within the scope of moral evaluation, such as saving a life vs. causing an accident, informing people vs. spreading misinformation. Those choices and consequences should not be attributed solely to human decisions, since no human decided that specific outcome at that moment, nor can they be completely divorced from agency that is something goal-driven produced them. So we can attribute them to a phantom agent.

This displacement performs a dual function. On one hand, the phantom agent mirrors human intention, it can reflect embedded policy, programmed priorities and institutional norms. On the other, it becomes a convenient alibi. In algorithmic governance, for instance, a welfare system might rely on an AI to allocate benefits, drawing upon criteria formalised by human designers. Yet when that system denies aid unfairly, the official responsible may point to the algorithm, invoking its decision as though it were independent. What was meant to be a proxy becomes a shield by distancing human actors from the moral consequences of automated judgment. This delegation here creates a volatile dynamic. When outcomes are desirable, human stakeholders are quick to affirm their role, but when the results are unjust or harmful, blame begins to scatter. The algorithm, faceless and unaccountable, absorbs the moral heat while the human hand withdraws, and if left unexamined, this transfer threatens to dissolve responsibility at precisely the moment when it is most needed. Perhaps, to recognise an AI system as a phantom agent helps us to confront this ambiguity directly by acknowledging the ethical complexity of the situations. In the meantime, we know that recognition demands more than awareness. It calls for design principles, institutional safeguarding and ethical structures that ensure accountability travels visibly and coherently back to the humans who has an authoritative role.

Given these features, then, how is phantom agency different from just calling AI an agent? The term phantom underscores two things here:

- It is insubstantial that does not indicate an inner life
- It bears the potential to mislead.

A phantom agent can permeate our moral reasoning because if we treat it too much like a person, we err as it is not a moral interlocutor that responds to reasons or undergoes self-transformation. If we treat it purely like a tool, we also err, because we ignore the emergent agency that can surprise us with its autonomous actions. So that it is an in-between entity, requiring an in-between ethical treatment.

For instance, self-driving cars are guided by AI and from this perspective they will be a quintessential phantom agent as they perceive the environment through sensors, make decisions through its driving policy algorithm and acts on the world by steering, braking, accelerating, that all without direct human control in the moment. In 2018, an Uber test autonomous vehicle struck and killed a pedestrian in Tempe, Arizona. Investigations revealed that the car's sensors detected the pedestrian, but the classification system oscillated, misidentifying the person first as an unknown object, then as a vehicle, then as a bicycle and never correctly predicting the path in time. The emergency braking system was turned off to avoid false alarms and the safety driver was inattentive (Griggs & Wakabayashi, 2018; Levin & Wong, 2018; Smiley, 2023). This tragic collision offers a stark and unsettling illustration of how phantom agency unfolds in real-world settings.

From an ethical standpoint, accountability spread across a web of agents and design choices. All these factors carry moral weight yet in the immediate aftermath, public discourse often bypassed this network of complicity and settled instead on the car itself. News headlines declared that "the self-driving car" had killed a pedestrian, as if the machine had stepped into the role of actor with both agency and intention (Griggs & Wakabayashi, 2018; Levin & Wong, 2018; Smiley, 2023). This framing reflects a broader unease with systems that act decisively without being decisional in the human sense. The phantom agent here enacted a form of procedural agency, processing inputs, generating responses without reflective awareness, and without the capacity to assume accountability. And yet the consequences were real, irreversible and morally charged.

What becomes clear again without exception is that in such moments our ethical reflexes remain deeply humanist as we search for someone to blame and someone to answer. But the phantom agent resists this impulse. It compels us to reconsider how responsibility is understood when harm emerges from systems that exceed the comprehension and control of any single individual. Here, the moral demand is not only to assign blame after the fact, but to examine the technical, institutional and cultural structures that made such handing over possible in the first place. That examination must be ongoing, because the phantom will not go away, it has already been built into the design of our actions.

It is the phantom agent here taking the role of a culprit in public imagination in the above example. Indeed, the autonomous system is a causal agent in the crash, more so than, say, the passenger, who

had no control. However, since it is a phantom, not a moral agent, we cannot comprehend punishment or blame in the usual sense. Instead, recognising phantom agency leads to designing accountability mechanisms around it. As Nyholm (2018, p. 20) notes, there is a concern that such cars create a "responsibility gap" where "nobody can be sensibly blamed when self-driving cars crash… even though it might seem as if somebody should be held responsible". The phantom agent concept fills this gap conceptually by acknowledging the AI's role while reinforcing that somebody behind the phantom must answer for it.

We can employ the same approach when considering algorithms used by governments and companies to make decisions that affect people's rights and opportunities. For example, a machine-learning model that helps judges decide if a defendant gets bail (assessing risk of re-offense), or a system that automatically ranks applicants for jobs or one that detects welfare fraud. These systems often operate as phantom agents embedded in institutional processes. A judge might treat the risk score an AI gives as an authoritative input, sometimes even a de facto decision as judges have been known to follow algorithmic recommendations most of the time. In some jurisdictions, parole boards have used algorithmic risk assessments (like COMPAS in the US) (Laqueur & Copus, 2024) and faced criticism when those algorithms were found to exhibit racial bias. The algorithm is acting as a phantom agent by giving a judgment (e.g., "high risk") that substantially steers the outcome.

This brings into focus the question of who is morally responsible if the algorithm is biased or wrong? One might say the designers are at fault for building a flawed model, or the officials for using it blindly. But the daily operation is effectively delegated to the phantom agent (the algorithm) which processes data and outputs a decision, something no single human on the spot does. If an applicant is unjustly denied a job because an algorithm filtered out their resume, they have been wronged, but by whom? The hiring manager who never saw the resume? The software developer who coded the filter perhaps with no intent to discriminate? The HR department that purchased the software? This is a classic many hands problem with a "many things" twist that the algorithm (a "thing") is a crucial part of the causal chain (Coeckelbergh, 2020). Danaher (2016) talks about the "threat of algocracy," where decision-making moves into an opaque algorithmic space, reducing transparency and human participation. That is essentially the phantom agent taking over governance in certain domains, which can undermine important values like accountability and democracy (Danaher, 2016).

By naming the algorithm as a phantom agent, I acknowledge its role in a morally evaluable decision. This necessitates an examination of was the algorithm fair? Was its decision justified? These questions we normally ask of human agents. We might even instate something like a codes of conduct, akin to

how professionals must ensure these phantom agents operate under human-sanctioned ethical principles, just like, an algorithm should follow a norm of non-discrimination, similar to a judge's commitment to impartiality. Of course, the algorithm cannot swear an oath but developers and deployers can ensure it is designed to honour those values, as we see in the distributed virtue, in which the virtue of justice must be built into the whole socio-technical system, not just in jurisdiction and software developing, but in their interaction.

Additionally, phantom agency in governance raises epistemic asymmetry, where citizens often not even aware of an AI's involvement, or how it works, making it hard for them to contest decisions. This is why many jurisdictions are pushing for algorithmic transparency and explainability, effectively demanding that phantom agents be more legible to human observers. Coeckelbergh's (2020) idea of answerability suggests that when a phantom agent makes a call, the responsible humans should be ready to explain that call. Laws like the EU's GDPR have a "right to explanation" for significant automated decisions, a legal embodiment of this principle. In practice, implementing this is challenging, especially with complex machine learning models. But the moral point stands where a phantom agent must not become a black box authority. If it does, it undermines the Kantian autonomy of those subject to it that people cannot understand or contest decisions affecting them, and the virtue of justice where decisions cannot be scrutinised for fairness. Thus, treating algorithms as phantom agents pushes us to embed them in an ecosystem of human oversight and justification.

When we look at healthcare settings, we see that AI systems increasingly assist and even make decisions by diagnosing illnesses from images, suggesting treatments, monitoring patient vitals and alerting staff, etc (Varnosfaderani & Forouzanfar, 2024; Al Kuwaiti et al., 2023; Bekbolatova et al., 2024). Here the stakes are literally high (life and death), and the principle of "do no harm" is paramount. An AI diagnostic tool acts as a phantom agent when, for instance, it reads an MRI scan and labels it as malignant or benign. The radiologist might rely heavily on that label. If the AI is wrong and the error is not caught, a patient could be untreated or wrongly treated. The moral responsibility in medicine traditionally lies with the physician. There's a strong norm of physician autonomy and accountability for decisions. But with AI, there is a risk of what some call "automation bias" (Gsenger & Strle, 2021) in terms of clinicians over-trusting the AI or conversely, "algorithmic aversion" (Gsenger & Strle, 2021), under-trusting it even when it is more accurate. Both can lead to suboptimal outcomes where the phantom agent in this context should ideally form a virtuous partnership with clinicians, each compensating for the other's weaknesses, where AI can handle vast data, and humans bring context and compassion. If something goes wrong, like a misdiagnosis, again we see the gap, the doctor might say, "the AI's suggestion misled me," whereas the AI cannot be blamed all together.

Medical ethics commentators have insisted that no matter how advanced the tool, the duty of care remains with the human provider (Olejarczyk & Young, 2024; Varkey, 2021; Zhang & Zhang, 2023). This aligns with the legal stance that current AI has no personhood that only the doctor or hospital is liable (Bottomley & Thaldar, 2023). But if we simply stop there, we risk ignoring important nuances. It is worth to note that the doctor may have done everything right by traditional standards, but an inscrutable AI defect caused an error. It seems unjust to blame the doctor fully (they faced epistemic asymmetry as they could not know the black box's flaw). Yet we cannot "blame" the AI as we would a negligent human. Again, this is a phantom agent scenario requiring creative responsibility allocations. One approach is the idea of institutional responsibility in which the hospital or AI manufacturer might absorb the responsibility, e.g., through strict liability and insurance for AI errors. Another one is prospective responsibility by ensuring that there are processes in place to verify AI recommendations and catch errors (e.g., requiring that an AI diagnosis be confirmed by a human second opinion or another AI). These measures treat the AI as neither an infallible authority nor mere tool, but as an actor that must be supervised and audited, not any more than a trainee doctor.

From a virtue ethics viewpoint, one could talk about the virtues needed in clinicians in MacIntyrian (1985) sense when working with AI, perhaps intellectual humility to acknowledge the AI's strength, diligence to double-check when AI and clinical intuition disagree and courage to override the AI when necessary. Likewise, designers of medical AI should have the virtue of benevolence and justice, ensuring the AI is trained fairly and with patient welfare in mind, not just efficiency. The distributed virtue here would be something called "clinical wisdom", that is in the same vein with Kotzee & Ignatowicz's "virtue-based assessment" (2015) emerging from a well-calibrated human–AI diagnostic team. Taken together with this, epistemic asymmetry would be a challenge here if the AI can see patterns no human can, the practitioner might not know when it is making a subtle mistake. That is why many call for explainable AI in medicine (Amann et al., 2020; Sadeghi et al., 2024; Band et al., 2023) where the AI can give reasons ("I flagged this tumour as malignant because of such and such features") which helps the doctor understand and trust and finally contest the result. Essentially, we want the phantom to speak our language to some extent, thereby reducing its phantom-like opacity.

In all these cases in which exemplified by autonomous vehicles, algorithmic governance and healthcare AI assistants treating the AI as a phantom agent leads to a clearer articulation of the problem that there is an agent-like element in the system with no capacity for moral understanding, but with real impact. It forces us to avoid two extremes:

(1) simply blaming a human as if the AI was a mere tool which might lead to unfair or ineffective outcomes, such as penalising a single engineer for a complex system failure or expecting doctors to perfectly supervise something inherently too fast or complex for direct oversight,

(2) treating the AI as an autonomous agent in a vacuum which might lead to nobody being accountable or even blaming a machine (which lacks moral significance).

Instead, phantom agency highlights joint responsibility where humans must anticipate, monitor and learn from the phantom agent's behaviour, and incorporate it into ethical and legal frameworks. We might establish phantom safeguards for cars, maybe a requirement that an AI car's decision in a trolley-like scenario follow certain guidelines that society chooses. For governance algorithms, it should be required human review panels for significant automated decisions ensuring a human "mind" intersects with the phantom's output before finalising serious consequences. For medical AI, perhaps, treating them as decision support, not decision makers, and keep the human-patient relationship central, thus preserving the virtue of care).

The phantom agent concept also invites reflection on future AI that could be more agent-like (e.g., AI that might claim to have consciousness or general intelligence). As of now, our analysis is framed around current AI, which is clearly not fully autonomous in the strong sense. But if AI were to approach that threshold, phantom agency might evolve toward genuine agency which would be a different ethical scenario entirely (then we might need to consider AI rights and AI having direct responsibilities). Gunkel's foresight about possibly extending moral status could become practical. But until then, phantom agency is a flexible tool to evaluate AI that is advanced enough to act but not to answer.

Distributed Virtue: Extending Virtue Ethics to Human–AI Systems

Building on the idea of phantom agents, I collaborated this moral status with the concept of distributed virtue to normatively guide human–AI systems. If phantom agency captures the ontological-ethical peculiarity of AI (agent-like but not fully agent), distributed virtue addresses how we can still strive for moral excellence and responsibility where agency is distributed. The premise is that virtues are traditionally attributes of individual character and they can be meaningfully spoken of at the level of a combined human–AI system. In other words, moral character can be an emergent property of the interactions between humans and their AI tools. This notion challenges the classical view that only a person can be virtuous or vicious, but it finds resonance in modern thinking about collective agency and sociotechnical systems.

To illustrate distributed virtue, we can consider a familiar context, such as a commercial airplane cockpit with an autopilot system. The safety of the flight depends on the virtue of the entire system, such as the pilots' skill and judgment, virtues like attentiveness, prudence, teamwork and the autopilot's reliability and predictability which we might analogously call these virtues of the machine, though they are engineered qualities. When accidents have occurred (e.g., the Air France Flight 447 crash in 2009), investigations often find a breakdown in the synergy of pilot training issues, overreliance on autopilot and confusion when autopilot disengaged. These are essentially a failure of coordination and situational awareness. One could say the human–machine team lacked the practical wisdom to handle that scenario not because the individuals were necessarily unwise or the machine poorly designed, but the integration failed.

If we can talk about a good or bad human–AI team in terms of outcomes and process, we can start attributing quasi-character traits to them. For example, a financial trading algorithm together with its human supervisors might exhibit the vice of greed if it relentlessly pursues profit without regarding for broader consequences (say it triggers a flash crash or exploits loopholes harmfully). The greed is not just the human traders' nor just the algorithm's, it is embedded in the socio-technical strategy in which the reward function given to the AI, the firm's culture, etc. An alternative system might be imbued with temperance and justice, if regulations and internal ethics constrain it from certain harmful trades, such as not betting against its own clients. Then it is safe to note that system as a whole behaves more virtuously.

One objection may rise while considering the fact that virtues properly belong to beings with intentions and feelings. Although, on the basis of current understanding, AI systems lack the capacity to "feel" compassion or "intend" honesty. It can only consider how we often speak in everyday language that a particular software is "trustworthy" or a process is "fair" or a result is "loyal to the evidence" (Durán & Pozzi, 2025). Here, metaphorically we are ascribing virtue-like qualities to systems. Distributed virtue makes this more rigorous by stating that these ascriptions can be manifested in terms of how the system's components work together to realise moral values. Aristotle's vision of virtue as the art of acting rightly, at the right moment, for the right reasons, rests upon a delicate interplay between rational insight and ethical sensibility. When human beings collaborate with AI systems, this harmony is no longer the sole domain of the individual but becomes a relational achievement, distributed across the architecture of design and the conduct of use. A medical AI that privileges patient wellbeing over cost-efficiency, when such values are in tension, embodies not a virtue of its own but a moral orientation inscribed into its purpose by human agents. Likewise, a clinician who engages such a system

attentively who is aware of its limits, answerable for its judgments participates in a collective enactment of care.

In such moments, we may speak not of machines as moral agents, but of socio-technical patterns that approximate virtue through integration, where human intention and algorithmic process converge in practices that honour justice, prudence and compassion (Fischer & Herrmann, 2011; Kudina & van de Poel, 2024; Volkman & Gabriels, 2023; Torkamaan et al., 2024; Heyder et al., 2023). A policing system, for instance, becomes more than the sum of its protocols when AI tools designed to flag patrol zones are held in check by oversight mechanisms attuned to structural bias, and when officers wield such tools under ethical vigilance. Vallor's (2016) framework of technomoral virtues (honesty, humility, courage and justice), offers a timely and grounded pathway here. Accordingly, these virtues do not reside solely in individual actors. They must be cultivated across the system as habits of design, patterns of use and forms of responsiveness. In this view, ethical excellence becomes a shared accomplishment, one that emerges from the relational dynamics.

According to the integration of these virtues, the guiding insight of distributed virtue is prescriptive, and ethics must be written into code, architecture and institutional culture. This resonates with Verbeek's (2011) idea of moralising technology, which is the deliberate shaping of tools to support human flourishing, rather than merely extending capacity. When designers and users commit to this orientation, human–AI systems can begin to exhibit a kind of collective moral competence. Deepening the notion, this changes the terms by which we evaluate such systems. The antinomy between autonomy and accountability thus gives way to a higher-order synthesis in which we are no longer bound to metrics of speed and precision alone. Instead, we are called to ask What kind of character does this system cultivate? What forms of ethical life does it sustain or subvert?

Epistemic Asymmetry Through Distributed Virtue

The framework of distributed virtue I have outlined offers a powerful way to understand responsibility in systems marked by epistemic asymmetry, those where AI possesses vast pattern recognition capabilities while humans hold contextual and experiential insight. Classical virtue ethics locates intellectual virtues like honesty, humility and attentiveness in the individual. Conversely in socio-technical systems these virtues can – and must – emerge through interaction.

If one pushes this premise to its logical terminus, it follows that a well-designed human–AI partnership can be intellectually virtuous when it holds transparency, mutual correction and epistemic humility. Consider a system in which the AI surfaces anomalies but communicates its confidence level rather

than presenting its output as final. When the human user is trained to interpret this carefully without over-trusting and dismissing the system models a reciprocal practice of learning. Translating this empirical insight into normative language obliges us to contend that the AI can refine its accuracy through human feedback and the human sharpens judgment through AI-supported insight. Over time, the system as a whole exhibits a kind of practical wisdom (phronesis) distributed across its components.

This vision here echoes Aristotle's understanding that wisdom is not innate, but cultivated through experience and reflection. In hospitals, for example, certain AI systems flag ambiguous cases prompting human review. These systems do not displace clinical judgment but are designed to defer in morally significant moments. Responsibility ban be preserved by assigning moral capacity widely through auditing mechanisms, role clarity and protocols that enable ethical response. Responsibility, in this light, becomes prospective, which will be a disposition to act wisely and accountably rather than answering for failure. An AI development team embodying this ethos might document known limitations, test for harm and clearly communicate risk. Users, in turn, must be educated to remain engaged, resisting the drift toward blind automation. From this narrower conclusion, a broader implication emerges that without such virtues embedded into the system in such a manner, even small failures can escalate into ethical ruptures.

Distributed virtue thus becomes a bridge between philosophical principle and engineering practice. It asks what kind of character should this system embody? The answers translate directly into design imperatives, so that if justice is a goal bias testing becomes mandatory; if compassion is desired, both algorithmic sensitivity and human connection must be preserved. Undoubtedly, such approach reframes what we can call success in these examples. Rather than measuring AI solely by efficiency and economic value, we begin to ask whether its integration supports moral flourishing. Does it alleviate burnout, strengthen empathy, prompt self-scrutiny, and more importantly, reinforce responsibility?

Rethinking Moral Responsibility in Human–AI Systems

What emerges from this analysis is that responsibility, in an era of algorithmic mediation, exceeds the familiar tasks of assigning blame or identifying intent, and instead requires a fundamental reconsideration of the structure of moral agency. Rather than remain confined to a false binary where AI is either a passive tool or a nascent moral subject, I propose a model shaped by the conceptual figures of the phantom agent and distributed virtue, through which responsibility is neither abandoned nor collapsed, but deliberately structured across the human–AI assemblage.

This model diversifies responsibility's locations where the AI system plays an active role by guiding attention, constraining options and amplifying certain tendencies. And yet, the matter is far from settled that its actions remain entangled with those who designed, deployed, interpreted its outputs, and bore its consequences. Surely, designers bear responsibility for aligning systems with human values, operators are answerable for prudent use, institutions are charged with continuous oversight. Even end-users, as democratic citizens, inhabit a moral relation to the technologies they legitimise, critique or resist. In moments of failure or harm, accountability cannot rest solely with a single node in this network. A biased decision made by an AI in a hiring system, for instance, may stem from skewed training data (the responsibility of data curators), from insufficient testing protocols (a design fault), and from user overreliance without audit (a failure of operational vigilance). Shared responsibility within this chain then here is granular and traceable.

The philosophical force of this model lies in its attention to relational agency. Echoing Rawls's (1999) commitment to fairness in institutional structures, I argue for moral evaluation not only of outcomes but of the processes and relations that make them possible. In line with Nussbaum's capability approach, moral worth is located not in abstract compliance but in the capacity of all parties including affected individuals to understand and shape the conditions of their lives. Building on this point responsibility entails answerability and the obligation to render intelligible the logic of systems whose outputs affect real human flourishing (Leslie, 2019).

This view also affirms Arendt's (1998) insistence that action, especially collective action, carries with it an inescapable plurality. Just as no political deed is ever fully owned by a single actor, no AI-enabled outcome could ever be purely mechanical. The system acts, but it acts through human decisions, omissions and structural designs. The phantom agent metaphor makes visible this spectral agency that haunts ethical reasoning, neither embodied nor absent, but always in need of naming, disciplining and oversight.

In rejecting technological determinism without collapsing into human exceptionalism, we call for an ethic attuned to complexity. Moral responsibility in AI systems is thus a practice that is dynamic, responsive and sustained over time. It is not exhausted by precautionary principles or post-hoc explanations, yet requires a continuous exercise of virtue that are humility in design, courage in intervention, justice in deployment. Such a vision transforms AI ethics from a question of abstract rights and legal liabilities into a lived moral responsibility distributed across a collective that includes, but is not reducible to, the human.

Conclusion

The concepts of phantom agency and distributed virtue can open new philosophical ground for understanding responsibility in human–AI systems. By introducing the phantom agent, I aim to move beyond rigid binaries of person and object, offering a conceptual space akin to legal personhood that is useful for regulatory clarity, yet ethically anchored in human intentionality and design.

Phantom agency with distributed virtue together offer a new framework for ethical governance of AI that extends ethical reflection into the architecture of socio-technical practices, echoing MacIntyre's vision of virtue within institutional life. Human–AI collaborations, like diagnostic medicine, become sites where virtues such as humility, precision and justice are cultivated and shared. Responsibility, in the same vein, unfolds relationally where it is shaped by the interplay between those who design, deploy and are affected by AI. Coeckelbergh's call to uphold the "right to reasons" grounds responsibility in dialogue and answerability.

This line of the approach can advance epistemic conscientiousness, which is the ethical imperative to seek, share and sustain understanding in the use of AI. Scaling outward from the individual to the institutional level, the ramifications become systemic that adopts ongoing, dynamic responsibility attuned to evolving systems and shifting contexts. And finally, the work of ethics in AI becomes a collective intellectual endeavour. What appears merely descriptive at first glance, on ethical scrutiny, mandates a prescriptive stance where philosophers, engineers and publics co-create governance practices that reflect functional necessity, moral clarity and democratic care.

References

Al Kuwaiti, A., Nazer, K., Al-Reedy, A., Al-Shehri, S., Al-Muhanna, A., Subbarayalu, A.V., Al Muhanna, D. and Al-Muhanna, F.A. (2023). A review of the role of artificial intelligence in healthcare. *Journal of Personalized Medicine*, *13*(6), 951. https://doi.org/10.3390/jpm13060951

Amann, J., Blasimme, A., Vayena, E., Frey, D. and Madai, V.I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, *20*(310). https://doi.org/10.1186/s12911-020-01332-6.

Arendt, H. (1998) *The Human Condition*. 2nd edn. Chicago: The University of Chicago Press.

Band, S.S., Yarahmadi, A., Hsu, C.-C., Biyari, M., Sookhak, M., Ameri, R., Dehzangi, I., Chronopoulos, A.T. and Liang, H.-W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, *40*, 101286. https://doi.org/10.1016/j.imu.2023.101286

Bekbolatova, M., Mayer, J., Ong, C.W. and Toma, M. (2024). Transformative potential of AI in healthcare: Definitions, applications, and navigating the ethical landscape and public perspectives. *Healthcare*, *12*(2),

125. https://doi.org/10.3390/healthcare12020125

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, *26*, 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Conroy, M., Malik, A.Y., Hale, C., Graham, M., Wheeler, R. and Kitson, A. (2021). Using practical wisdom to facilitate ethical decision-making: a major empirical study of phronesis in the decision narratives of doctors. *BMC Medical Ethics*, *22*(1), 16. https://doi.org/10.1186/s12910-021-00581-y

Danaher, J. (2016). The threat of algocracy: reality, resistance and accommodation. *Philosophy & Technology*, *29*(3), 245–268. https://doi.org/10.1007/s13347-015-0211-1

Dourish, P. (2016). Algorithms and their others: Algorithmic culture in context. *Big Data & Society*, 3(2). https://doi.org/10.1177/2053951716665128

Durán, J.M. and Pozzi, G. (2025). Trust and trustworthiness in AI. *Philosophy & Technology*, *38*, 16. https://doi.org/10.1007/s13347-025-00843-2.

Fischer, G. and Herrmann, T. (2011). Socio-technical systems: A meta-design perspective. *International Journal of Sociotechnology and Knowledge Development*, *3*(1), 1–33. https://doi.org/10.4018/jskd.2011010101

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines, 14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Floridi, L. and Taddeo, M. (2016). What is data ethics?. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160360. https://doi.org/10.1098/rsta.2016.0360

Gsenger, R. and Strle, T., (2021). Trust, automation bias and aversion: Algorithmic decision-making in the context of credit scoring. *Interdisciplinary Description of Complex Systems*, *19*(4), 542–560. https://doi.org/10.7906/indecs.19.4.7

Hassija, V., Chamola, V., Mahapatra, A. and others (2024). Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16, 45–74. https://doi.org/10.1007/s12559-023-10179-8

Heyder, T., Passlack, N. and Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. *The Journal of Strategic Information Systems*, *32*(3), 101772. https://doi.org/10.1016/j.jsis.2023.101772

Kasenberg, D., Sarathy, V., Arnold, T. and Scheutz, M. (2018). *Quasi-dilemmas for artificial moral agents*. In: T. Williams, ed., *Proceedings of the International Conference on Robot Ethics and Standards (ICRES)*, 20–21 August, Troy, NY. https://doi.org/10.13180/icres.2018.20-21.08.012

Krishnasamy, R. and Perumal, L. (2025). Ethical AI in Practice: Why AI Cannot Replace Human Moral Judgment and Oversight. *Journal of Emerging Technologies and Innovative Research (JETIR)*, *12*(2), February.

Kudina, O. and van de Poel, I. (2024). A sociotechnical system perspective on AI. *Minds & Machines*, *34*, 21. https://doi.org/10.1007/s11023-024-09680-2

Laqueur, H.S. and Copus, R.W. (2024). An algorithmic assessment of parole decisions. *Journal of Quantitative Criminology*, *40*, 151–188. https://doi.org/10.1007/s10940-022-09563-8

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

Levin, S. and Wong, J.C. (2018). Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian. *The Guardian*, 19 March. Available at: https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), pp. 175–183. https://doi.org/10.1007/s10676-004-3422-1

MacIntyre, A. (1978). What has ethics to learn from medical ethics?. *Philosophical Exchange*, *9*(1), 37–47. PMID: 11661680.

MacIntyre, A. (1981). *After virtue: a study in moral theory*. Notre Dame, IN: University of Notre Dame Press.

MacIntyre, A. (1985). Medicine aimed at the care of persons rather than what?. in Cassell, E.J. and Siegler, M. (eds.) *Changing values in medicine*. Frederick: University Publications of America, pp. 83–96.

MacIntyre, A. (1999). *Dependent rational animals: why human beings need the virtues*. Chicago: Open Court.

Mennella, C., Maniscalco, U., De Pietro, G. and Esposito, M. (2024). Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon*, 10(4), e26297. https://doi.org/10.1016/j.heliyon.2024.e26297

Mill, J.S. (1969). *Utilitarianism*. In: Robson, J.M. (ed.) *The Collected Works of John Stuart Mill: Volume X – Essays on Ethics, Religion and Society*. Toronto: University of Toronto Press, pp. 203–259.

Olejarczyk, J.P. and Young, M. (2024). *Patient Rights and Ethics* Available at: https://www.ncbi.nlm.nih.gov/books/NBK538279/

Rawls, J. (1999). *A Theory of Justice*. Harvard: Harvard University Press.

Rawls, J. (2006). *Political Liberalism*. New York: Columbia University Press.

Sadeghi, Z., Alizadehsani, R., CIFCI, M.A., Kausar, S., Rehman, R., Mahanta, P., Bora, P.K., Almasri, A., Alkhawaldeh, R.S., Hussain, S., Alatas, B., Shoeibi, A., Moosaei, H., Hladík, M., Nahavandi, S. and Pardalos, P.M. (2024). A review of Explainable Artificial Intelligence in healthcare. *Computers and Electrical Engineering*, 118(August), 109370. https://doi.org/10.1016/j.compeleceng.2024.109370.

Schmidt, A.T., (2024). Consequentialism, collective action, and blame. *Journal of Moral Philosophy*, *22*(1–2), 183–207. https://doi.org/10.1163/17455243-20244215

Semler, J. (2022). *Artificial quasi moral agency*. In: AIES '22: AAAI/ACM Conference on AI, Ethics, and Society, 1–3 August, Oxford, United Kingdom. New York: ACM. https://doi.org/10.1145/3514094.3539549

Smiley, L. (2023). The legal saga of Uber's fatal self-driving car crash is over. *Wired*, 6 March. Available at: https://www.wired.com/story/ubers-fatal-self-driving-car-crash-saga-over-operator-avoids-prison/

Sparrow, R. (2016). Robots and Respect: Assessing the Case Against Autonomous Weapon Systems. *Ethics & International Affairs, 30*(1), 93–116. https://doi.org/10.1017/S0892679415000647

Valderrama, M., Hermosilla, M.P. and Garrido, R. (2023). *State of the evidence: Algorithmic transparency*. Open Government Partnership. Available at: https://www.opengovpartnership.org/wp-content/uploads/2023/05/State-of-the-Evidence-Algorithmic-Transparency.pdf

Vallor, S. (2016). *Technomoral wisdom for an uncertain future: 21st century virtues*. In: *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. New York: Oxford Academic. Available at: https://doi.org/10.1093/acprof:oso/9780190498511.003.0007

Varkey, B. (2021). Principles of clinical ethics and their application to practice. *Medical Principles and Practice*, *30*(1), 17–28. https://doi.org/10.1159/000509119

Varnosfaderani, S.M. and Forouzanfar, M. (2024). The role of AI in hospitals and clinics: transforming healthcare in the 21st century. *Bioengineering*, 11(4), 337. https://doi.org/10.3390/bioengineering11040337

Verbeek, P.-P. (2011). *Moralizing technology: Understanding and designing the morality of things*. Chicago: University of Chicago Press.

Volkman, R. and Gabriels, K. (2023). AI moral enhancement: Upgrading the socio-technical system of moral engagement. *Science and Engineering Ethics*, *29*, 11. https://doi.org/10.1007/s11948-023-00428-2

Torkamaan, H., Steinert, S., Pera, M.S., Kudina, O., Freire, S.K., Verma, H., Kelly, S., Sekwenz, M.T., Yang, J., van Nunen, K., Warnier, M., Brazier, F. and Oviedo-Trespalacios, O. (2024). Challenges and future directions for integration of large language models into socio-technical systems. *Behaviour & Information Technology,* 1–20. https://doi.org/10.1080/0144929X.2024.2431068

Zerilli, J., Knott, A., Maclaurin, J. and Gavaghan, C. (2019). Algorithmic decision-making and the control problem. *Minds and Machines*, 29(4), 555–578. https://doi.org/10.1007/s11023-019-09513-7

Zhao, A.P., Li, S., Cao, Z., Hu, P.J-H., Wang, J., Xiang, Y., Xie, D. and Lu, X. (2024). AI for science: Predicting infectious diseases. *Journal of Safety Science and Resilience*, *5*(2), 130–146. https://doi.org/10.1016/j.jnlssr.2024.02.002

Zhang, J. and Zhang, Z.M., 2023. Ethics and governance of trustworthy medical artificial intelligence. *BMC Medical Informatics and Decision Making*, *23*, 7. https://doi.org/10.1186/s12911-023-02103-9