Philosophy of AI
Vol. 2, 2026, 129-144
DOI: 10.18716/ojs/phai/2026.3326

BEARING THE MORAL WEIGHT: WHAT'S LOST WHEN AI DECIDES

Christine Susienka[*]

Sacred Heart University, Philosophy Department, Fairfield, United States

[*]*csusienka@gmail.com*

Abstract This article argues that outsourcing high-stakes decisions to AI runs the risk of degrading human dignity, and that this is true regardless of whether or not a given AI system consistently arrives at the same conclusions as human experts. I focus specifically on cases where there are relational or role-related obligations already in place, such as those of judges, doctors, and citizens. Failing to offer human decisions in these contexts degrades both the dignity of those about whom the decisions are made and that of the human beings who outsourced the decisions. With regard to the former, it does so by denying them recognition respect, which requires what I characterize as 'bearing the moral weight.' With regard to the latter, it does so by undermining the agency and dignity that results from taking effective responsibility for oneself and for one's society. I conclude that for these dignity-related reasons, we ought to be leery of outsourcing human decision-making in high-stakes cases regardless of the progress we are able to make on primarily technical ethical challenges like bias, explanation, and accountability.

Keywords Human dignity; moral agency; vulnerability; recognition respect; relational responsibilities

1. Introduction

That artificial intelligence (AI)[1] poses ethical challenges is at this point well-known. Issues of bias embedded in algorithms, AI hallucination, and the 'black box' problem have shown up in several high-profile cases over the last few years (e.g., Babic et al., 2021; Metz & Weise, 2025; Obermeyer et al., 2019). Nonetheless, at least among techno-optimists, there remains confidence that these are growing pains, and that it is both possible and desirable for us to develop and use AI in decision-making because (at least in part) doing so is more objective (e.g., Gow, 2022; Victoroff, 2022). Even if our tech isn't there yet, the argument goes, ultimately AI will be better able to treat people fairly than we humans are likely to do left to our own devices. Some might go so far as to wonder if we have a responsibility to outsource

1. How we should define AI or draw the boundaries around what counts as AI is contested. For the purposes of this article, little hinges on the definition, though the concerns that I raise apply primarily to generative and predictive AI rather than to classical symbolic AI.

some kinds of decision-making on account of our human fallibility. Perhaps, in at least some contexts, we do one another a disservice by not turning to the machine.

Nonetheless, additional scholarship raises incisive questions about whether it is even possible for us to develop AI that is objective in the sense of being value-neutral and free of morally problematic biases (e.g., Crawford, 2021). It is, of course, human decision-makers who program algorithms, select training data, code that data, and determine the contexts in which AI is implemented. For these reasons, eliminating human biases and fallibility from the process is not possible. However, the distance between the humans engaged in these processes and the end users who see the output obscures the many hands involved and gives the illusion of a kind of objectivity that fails to reflect the reality.

One might still argue that this, too, is primarily a problem of human fallibility. The newness of this technology and the average person's unfamiliarity with its limitations is a problem of the moment. With sufficient human supervision and AI education, we can dispel the illusion that AI results could become fully objective. A weaker version of the goal, then, might be that ethically programmed AI systems with sufficient human supervision will be at least as good, and often better, at avoiding mistakes and biases than human decision-makers alone. If this is possible, then, given the additional efficiency benefits of AI, there seem to be good reasons for embracing its use. This might be even more true in the contexts of high-stakes cases.

Nonetheless, high-stakes cases warrant heightened attention to potential ethical implications. This is so because the moral stakes are also heightened. While there is a robust literature focusing on autonomous weapons systems (AWS) and autonomous vehicles (AV) as instances of high-stakes decision-making, there is less about high-stakes cases that are part of the backdrop of everyday civic life and that reflect agents in positions of authority fulfilling (or failing to fulfill) duties of care that they possess in virtue of their particular relationships to the individuals about whom high-stakes decisions are being made. Beyond more general moral duties (and the unique context of just war theory), I focus on duties such as those that a doctor has to their patients, those that a judge has to defendants and victims, and those that citizens have to one another. While these contexts are not always high stakes, they frequently enough are to deserve consideration.

While these cases bear similarities to AWS and AV, the decision-making-contexts that are my primary focus are importantly distinct. While AWS and AV contexts are often time-sensitive and emotionally charged in ways that inhibit on-the-ground human agents from being able to effectively retake control, the cases I focus on are not time-sensitive in the same way. Further, they concern duties of care and responsibility that stem from relationships that exist between the relevant parties. For these reasons, separate inquiry is warranted. However, there is much to be learned from the AWS and AV literature, and at points in this article I draw on insights from it, particularly with regard to the distinctiveness of moral judgments (especially Purves et al., 2015; Sparrow, 2021) and the importance of the attitudes communicated toward those about whom decisions are being made (especially Harris, 2020; JafariNaimi, 2018).

More specifically, this article argues that even if we could achieve accurate and unbiased results from AI, we must consider what is lost when we outsource high-stakes decision-making of this type to non-human intelligences. I maintain that the recurrent moral gap[2] present in these cases should be characterized as an absence of recognition respect (here I follow Darwall, 1977), as well as insufficient regard for the significance of shared vulnerability. Recognition respect requires that someone bear the moral weight of decisions that cut to the core of human experience. When someone is sentenced to life in prison or denied cov-

erage for a life-saving treatment, they deserve someone who appreciates the stakes for their particular human life making the decision and communicating it.

Part of the worry here is about ensuring that an identifiable agent can be held accountable for mistakes, but my concern goes deeper. To put it more colloquially, it is the idea that we have a responsibility for one another, and that the more we turn to artificial intelligence to arrive at life-altering decisions, the more that we outsource this duty. In doing so, I argue that we run the risk of degrading a more basic respect for human dignity, both our own as decision-makers and that of the individuals about whom decisions are made. This threat to human dignity would remain even if human experts were to agree that the decisions made by an AI were correct.

While the source of dignity is contested, the concept of dignity has done important work at keeping attention on the equal value and importance of each human being as, to use a Kantian turn of phrase, an end in themselves, worthy of both care and respect.[3] While human dignity is often taken to be inviolable, our treatment of one another can be degrading insofar as it signals a failure to recognize or take seriously the dignity of the other, and it can undermine our capacity to live lives worthy of our dignity.[4] In the case of AI-decision making in high-stakes cases, I argue that there is a failure of the former for the person about whom the decision is made, and a failure of the latter for the outsourcing decision-maker.

If I am right, this result is especially problematic in the current climate of decreased social and institutional trust. Rather than continue to improve ourselves and commit to genuinely seeing one another, outsourcing decision-making to AI suggests giving up on the potential for us to do so and on our responsibility to try. For these reasons, I ultimately argue that in at least some high-stakes cases, the contours of which ought to be determined through public conversation about values (following, e.g., Jasanoff, 2016; Sandel, 2012), there ought to be a right to a human decision (such as that advocated for by Tasioulas (2022) and Huq (2020)) and we should be especially leery about championing growth in, and the application of, AI in these areas without more clear-eyed public conversations about this possible loss.

To build my case, I first provide a brief overview of some of the ethically salient decisions made by human beings at various stages of AI development, refinement, and implementation. I do so to dispel the presumption that AI systems could ever be both accurate and value-neutral. In particular, efforts to arrive at one form of objectivity (accuracy, freedom from bias, or value-neutrality) still leave other moral problems intact and in need of human decision-making. This observation is important because objectivity is often presented as an ethical benefit of AI decisions as opposed to decisions made by human beings. Even with

2. The phrase 'moral gap' might call to mind responsibility gaps (a term coined in Matthias, 2004). I opt to not use that phrasing here because the primary contours of that literature focus on accountability in cases of an errant decision or a bad outcome and on diffuse responsibility in complex systems (see, e.g., Gunkel, 2020; Hevelke & Nida-Rümelin, 2015; Hindriks & Veluwenkamp, 2023; Sparrow, 2007). While I agree that these are serious concerns, my focus is on a moral gap that remains even if the decision is correct. Further, as I characterize it, the gap isn't about a failure of accountability for error; it's about a failure for specific moral agents to take up their responsibilities to particular other human agents.

3. For example, the concept of dignity was prominently invoked in international law and norms in the aftermath of WWII. The Universal Declaration of Human Rights (UN General Assembly, 1948) and the Basic Law for the Federal Republic of Germany (Parliamentary Council, 1949) are two clear examples. I take May and Daly (2019) to make a compelling case for why dignity is important and cannot be reduced to other values.

4. Here I follow Nussbaum (2007) in thinking of a life worthy of human dignity as one in which an agent is able to cultivate central capabilities that enable them to live a flourishing human life. For the purposes of this paper, one need not endorse Nussbaum's full set of capabilities. What is important is taking seriously that respect for dignity requires attention to both the attitudes that we communicate toward one another and how the actions that we take (individually or collectively) enhance or undermine the ability of another to cultivate central human abilities. We might also look to Lamberton et al. (2025). They argue that the complexity of dignity might make it a useful concept in AI conversations insofar as it can draw our attention to a range of important and related considerations that might otherwise go insufficiently attended to. For these reasons, I think an expansive rather than a narrow conception of human dignity is beneficial at this stage of our discussion on AI decision-making.

knowledge about the limitations of AI systems, there is still public speculation that the AI can be better than we are (see, e.g., Moschella, 2022), and there is a tendency for practitioners to be strongly influenced by its recommendations, even when they are generally aware of AI's limitations (see, e.g., Klingbeil et al., 2024; Messeri & Crockett, 2024; Vicente & Matute, 2023). For these reasons, it is important to emphasize AI's limitations. If I am right about this, then we ought to be cautious about normalizing presumptions that human experts should defer to the machine or that we ought not bother to become experts because AI can make high-stakes decisions at least as well as we can, if not better. Doing this work enables us to consider why there might still be a need for securing a right to a human decision for at least some high-stakes cases.

Next, I develop the positive argument that, in at least some high-stakes cases, we make a serious moral mistake when we outsource those duties to AI. My point is that morality requires not merely accountability for errors, but also an agent taking up responsibility for decisions and recognizing the weightiness of doing so. I characterize this lapse in responsibility in terms of a failure to bear the moral weight. When we fail to bear the moral weight, we fail to respect the dignity of the person about whom the decision is about by denying them recognition respect. I further argue that outsourcing moral judgments in high-stakes cases also degrades the dignity of decision-makers by reducing their capacities and effective responsibility. For these reasons, I conclude that there is a right to a human decision in at least some high-stakes decision-making contexts.

I end by suggesting that we ought to determine the contours of a right to a human decision through robust public deliberation which, in turn, demands that we actively take up effective responsibility for our societies and intentionally exercise our moral capacities.

2. The Illusion of Objectivity

One way of framing the problem of bias in AI is as a failure of accuracy. On this framing, a central problem with bias in AI models[5] is that they get the output wrong (see, e.g., Cerrato, Halamka, & Pencina 2022). They predict that people will default on loans who do not, or will reoffend who do not, and they do this at scale and in ways that reflect long-standing patterns of structural injustice. Approached in this way, the goal of eliminating bias in AI is straightforwardly aligned with the goal of making AI more accurate in its predictions. Eliminating bias is then both an ethical imperative and a requirement for ensuring improved accuracy.

Consider COMPAS, a risk assessment program used to predict recidivism rates by assigning scores to potential re-offenders (see, e.g., Crawford, 2021; Dressel & Farid, 2018; O'Neil, 2016). Because some of the data points used to make these predictions serve as proxies for race, black individuals have been disproportionately identified as high risk, and they have been given lengthier prison sentences than white individuals as a result. In addition, COMPAS's predictions have not proven to be especially accurate, and the kind of error it is more likely to make tracks racial identity. There have been more false positives when the potential re-offender was black and more false negatives when the potential re-offender was white. Especially given the long-standing problem of racial injustice in the US criminal justice system, this result is deeply disturbing as one of the purported benefits of COMPAS is that it

5. Given my focus on decision-making, one might wonder why I use the terms 'AI systems' and 'AI models' rather than 'AI agents.' While it is certainly true that many current AI are agents in the sense of being goal-directed and responsive to their environments, and that this term is regularly used in the literature, I worry that the frequent use of anthropomorphic terms obscures underlying ethical questions. See Deroy (2023) for a discussion of this problem in the context of the term 'trustworthy.' This point will be briefly discussed in Section 2.1.

provides a fair way of predicting risk. By removing the human factor from the end-stage of the process, we have created the illusion of increasing accuracy and fairness while the reality is anything but.[6] While COMPAS was not intended to be used as a stand-alone tool, in practice that is what it has frequently become. We should keep this lesson in mind as we think about incorporating AI systems with the expectation that they will be used as merely one piece of the decision-making puzzle.[7]

While the racial injustice resulting from the use of COMPAS would be a significant problem even if it were an outlier, COMPAS is far from alone. Morally problematic biases embedded into AI and then, in turn, used to perpetuate and justify historical injustices under the guise of objectivity via pernicious feedback loops have been identified in a wide range of contexts. These include, among others, job hiring practices (e.g., Laker, 2023; Venzke & Akselrod, 2023), medical assessments (e.g., Buslón et al., 2023; Obermeyer et al., 2019), and determinations of credit-worthiness (e.g., Garcia et al., 2024; Heaven, 2022).

One suggestion for alleviating this result is to significantly increase the size or diversity of training datasets (e.g., Kostick-Quenet et al., 2022; Wang et al., 2022). However, doing so does not necessarily reduce the kind of problem previously described if the pattern of injustice was widely practiced. In fact, the more wide-scale a structural injustice is, the more likely it is that an AI model will pick up on that pattern in the dataset. In a case like using AI to determine credit-worthiness, then, the problem is about more than just having a limited data set. It is also about labeling the data that you do have and coding the algorithms you are using in ways that mitigate bias. Doing so is not a purely technical feat, though it is that too. It also requires engaging in ethical reasoning and intentionally encoding values into a system that appears to (and is often presented to) outsiders as value-neutral (see, e.g., Silberg & Manyika, 2019).

Coding these algorithms, in turn, requires making decisions about morally significant cut-offs, about what factors should be marked as irrelevant, about whether any features should be given additional weight in the calculus, and so on. These are not value-neutral decisions, and they beg the question about the purpose of any given AI system. What has it been designed to recognize as success? What kinds of errors should be considered acceptable? These outcomes are often not straightforwardly aligned with the goal of increased accuracy. In the example of an AI system designed for use at banks to determine credit-worthiness, the determination might be made that getting it right is not the primary goal; the primary goal is minimizing false positives and in doing so avoid putting the bank in a financially risky position. The goal is not, therefore, to accurately assess risk; the goal is to mitigate risk by ruling out potentially risky borrowers even if some low-risk borrowers are ruled out too.

Thus, in attempting to mitigate the bias often embedded in and perpetuated by the widescale use of AI models, we need to do more than just increase our training data or code in ways that minimize and perhaps even counteract bias. We already need to be engaged in more robust conversations about what values we want to prioritize, how we ought to interpret them, and how transparently these decisions need to be communicated.

---

6. Equivant, the owner of COMPAS, rejects many of the critiques (Equivant Supervision, 2024). Much seems to hinge on the conception of fairness that is being operationalized.

7. COMPAS is far from the only case of a useful tool taking on a much greater role than it was designed for, and in doing so generating problematic results. The problem of p-hacking provides another example with wide-sweeping consequences. When discussing the role of p-values, Ronald A. Fisher was clear that they should not be the sole determinant of statistical significance. For a discussion of this topic, see Denworth (2019). In the next section, I follow up on this thread, arguing that while over-reliance in usage is not a conceptual problem, there are good reasons for believing over-reliance on an AI system would become a serious practical problem unless there are significant changes to present incentive structures.

Though sometimes the goal of reducing bias is to increase the accuracy of an AI model's output, the public flop of Google's Gemini AI Image Generator rollout in February 2024 (see, e.g., O'Brien, 2024) reflects an attempt to do something different, to proactively encourage the value of diversity. When responding to an inquiry, Gemini generated results that reflected additional hidden prompts that prioritized diversity. Google added these prompts to avoid the generation of offensive and stereotypical images, a long-standing problem with previous AI image generators and chatbots.[8]

If one were to search for scientists, Gemini was programmed to provide diversity in the racial and gender makeup of the proffered images. Without additional coding, an AI system drawing on a historical training dataset based on images of well-known scientists from media and history would likely have provided images of white men. In doing so, it would merely be reflecting prominent biases. However, Google's additional prompt, while perhaps well-intentioned, had the unfortunate side-effect of making glaring errors. For instance, when prompted to depict a group of people wearing the 1943 uniforms of Germans soldiers, Gemini generated a racially diverse group and refused to generate images of white men. Google Senior Vice President Prabkhakar Raghavan described what went wrong by saying, "Our tuning to ensure that Gemini showed a range of people failed to account for cases that should clearly *not* show a range" and "Over time, the model became way more cautious than we intended, and refused to answer certain prompts entirely—wrongly interpreting some very anodyne prompts as sensitive" (Raghavan, 2024).

The unexpected outputs in this case further reveal the extent to which explicit value-based coding decisions are constantly being made behind the scenes, and that aiming for accuracy and aiming to be bias-free are not always aligned objectives. The work of designers and programmers is about doing far more than merely ensuring accuracy. Accuracy is only one among many of the values prioritized either explicitly, as it was in the case of Gemini, or implicitly, perhaps without recognition of what the implications of a seemingly benign coding choice might be. Human hands and human choices are still all over many of the moving parts of AI results.

Often bias or embedded values in AI models only come to light after the AI has been in circulation for long enough that the pattern in output seems very unlikely to be a fluke, when there is sufficient outcry to motivate the company or institution using the model to more closely look into it, as occurred in the Gemini case. This mechanism for identifying mistakes has predictably not been good for public trust and confidence in usage. For this reason, one proposal is to increase transparency, to make AI models accessible and thus open to scrutiny.

While I am largely supportive of this move, what it would look like to successfully achieve this goal is not straightforward. In addition to the intellectual property defenses already being made (see, e.g., Hattingh, 2025), we might still wonder to whom the information is transparent. Unless we are also making a concerted effort to increase more general coding literacy, understanding of the models will still be attained by most people through others' interpretations, as independent watchdog organizations, government regulators, or concerned private parties provide their own reports on the accuracy and embedded values within a given AI model. And as the Gemini case reveals, unsupervised models can also end up generating outcomes that take intentionally embedded values to an extreme never intended by the original coders. Further, as AI models are constantly learning based on their

8.Microsoft's 2016 launch of Tay.AI is a prominent example. Within hours of the chatbot's launch, it was spouting hateful language and imagery. See Microsoft's response in Lee (2016).

usage, the assessment that a given AI model is accurate at achieving a given aim and is consistent with a given set of values would be a moving target.

In sum, values-based decisions made by human beings are infused throughout the process, and while we can make substantive progress on many of these issues, serious ethical questions remain about when and how we ought to deploy AI, especially for high-stakes decision-making.

3. Taking Responsibility and Bearing the Moral Weight

This section argues that outsourcing high-stakes decisions to AI, even if the AI system consistently arrives at the same conclusions as human experts, runs the risk of degrading human dignity. I consider this both in terms of the dignity of the person about whom the decision is made, and in terms of the dignity of the human decision-maker who has outsourced this responsibility. With regard to the person about whom the decision is made, I argue that not offering a human decision degrades the dignity of the person by denying them recognition respect. Even if an AI system can arrive at the same conclusion as human experts, it cannot offer genuine recognition. Second, I argue that an increasing pattern of outsourced moral responsibility undermines democratic relations and human connection by disincentivizing the attitudes and practices that lead citizens to proactively take effective responsibility for themselves and their societies. I maintain that this winnowing of agency undermines the dignity of the decision-makers themselves.

3.1 Moral Agency and Recognition Respect

In taking a particular person's situation sufficiently seriously, I maintain that we must not merely determine who is accountable when a mistake is made, but who should bear the weight of endorsing and executing a given judgment, even when that judgment is correct. Otherwise, we run the risk of turning moments of great moral salience into merely bureaucratic affairs; we perhaps even intentionally suppress our own individual judgment on the grounds that the AI knows best. In reporting an AI's conclusion, without taking it up and endorsing it as our own, we fail to adopt the stance of a moral agent. While an AI model might offer a probabilistic prediction about, say, a treatment plan for a given disease, or a proposed sentence based on a series of inputs, it cannot take up responsibility for the judgment to apply that recommendation to a given individual's case. Respect for human dignity requires that a human agent take up the further responsibility of endorsing *this* judgment for *this* person, with an awareness of what that means for that particular person's life.

In paradigmatic cases of responsibility gaps, the gap exists because it is unclear who ought to be held accountable or because it is unclear how we could meaningfully hold an AI accountable[9]. However, in cases like those that I have described, there is already an identifiable human agent who has been bearing particular responsibilities to given individuals. Doctors have duties of care toward their particular patients in virtue of their role as someone's doctor. Likewise, judges have particular role-based duties of justice and respect toward defendants[10] in their courtrooms. For these reasons, either replacing the decision-making of these human agents with AI, or heavily incentivizing doctors and judges to rely on AI in decision-making creates a somewhat different moral gap. It demeans the value of these roles

9. See FN 2 for a brief discussion of responsibility gaps.

10. Judges also have role-based duties of justice and respect toward victims. I here focus on defendants because the decisions in the courtroom are explicitly about defendants' sentences. However, these decisions do, of course, have serious implications for victims and their families as well.

and flattens the relationships that exist between doctors and patients and between judges and the public.

After all, holding someone accountable isn't only about giving them the correct praise or punishment; it's also about *us* holding them responsible, and about the relationship that exists between the parties. The dynamic of holding one another responsible is itself part of what it is for us to recognize one another as fellow moral agents. In the background of the intuition that I explore here is a commitment to recognition respect of the sort initially described by Stephen Darwall (1977), i.e., the kind of respect that people are entitled to simply on account of being human moral agents.

Recognition respect goes unmet when no one bears this moral weight, or when that weight is so diffuse within a community that no one acknowledges that they carry it; this moral responsibility is just a background condition of everyday life that never feels salient. This is especially problematic in cases like the judge and the defendant or like the doctor and their patient, cases where the relationships that they stand in generate specific role-based duties between them. Outsourcing high-stakes decision-making to AI systems deprives the defendant and the patient of the recognition respect that they deserve on account of their human dignity. Recognition requires being seen by the other, being taken seriously as a moral equal. While one might argue, for instance, that subjecting everyone to AI courtroom decisions would be one way of treating us as moral equals (assuming that we've stipulated success as the reduction of bias and increased transparency), I argue that this would still be a failure of recognition. Though it would treat us all the same, it would fail to create conditions of being recognized by the judge as a human being of equal moral worth in a position of relational vulnerability.

If no takes up the responsibility, then no one is bearing its weight, no one is engaging in the emotionally and cognitively complicated state of bearing witness and in looking in the face the consequences of this high-stakes choice. Even when we make the right decision, human persons making high-stakes decisions are still often freighted with the significance of their choices, and we might have good reason to express concern if they were not. Making a life-changing determination about another person ought not leave us cold, and so much the worse for us if it does. Making and communicating a decision is a process with moral stakes for everyone involved, including the decision-maker; reporting the output of an AI system fails to reach the threshold of moral responsibility for the agent reporting the decision and fails to show recognition respect for the person about whom the decision is made.[11]

One might here object that I have rigged the argument by presuming that AI systems are not agents. Of course, the language of AI agency has become increasingly widespread as AI systems have become capable of a wider range of complicated tasks. While I do not deny that many AI systems possess the rudimentary features of agents, for instance, they are goal-directed and responsive to inputs from their environment, I maintain that the kind of agency of which they are capable fails to rise to the level of proactively taking moral responsibility. An AI agent cannot take up their responsibilities, show recognition respect, or bear the moral weight of a decision. Even the rise of agentic AI, which is more adaptive, autonomous, and able to perform a wider range of tasks without additional human input (see, e.g., Acharya et al., 2025), provides no reason for presuming that AI agents will become capable of self-reflection and recognition.[12] And even then, even if they could bear the moral weight, there would be a further question about whether we humans ought to offload that

11. As a caveat, I am not here addressing whether a legal right to a human decision is already entailed by present laws in any specific legal context. I'm considering whether a right to a human decision is a moral right or a basic requirement of human dignity that warrants protection.

ethical responsibility to an agent who lacks the relational obligations that we bear to one another.

I am not alone in rejecting the claim that AI can rise to this level of complex, situated moral agency on the grounds that, even if we grant that AI can be agential, it is the wrong kind of agent for rising to this challenge. AI is disembodied (or embodied differently from human beings in ways that are morally salient), lacks self-consciousness, is emotion-free, and does not participate in the kind of shared vulnerability that can give rise to compassion. Self-consciousness, embodied experiences, and shared vulnerability are important factors in both enabling genuine moral judgment and in being able to authentically communicate recognition respect.

According to Sparrow (2021) "Ethics is personal in a way that science is not" (685) and "the force of an ethical claim depends in part on the life history of the person who is making it" (685). For Purves et al. (2015), "Moral judgment requires either the ability to engage in wide reflective equilibrium, the ability to perceive certain facts *as* moral considerations, moral imagination, or the ability to have moral experiences with a particular phenomenological character" (851-852). Purves et al. further maintain that AI[13] cannot in principle possess these capacities. Nassim JafariNaimi (2018) raises a similar point in the context of autonomous vehicles. She argues that "Ethical theories are key in helping us analyze and understand situations and imagine, devise, and assess plausible courses of action in response—but they are not recipes for action" (309). In particular, these thinkers all highlight moral imagination and moral agency as entailing a certain kind of phenomenological experience that is part of realizing one's duties and determining how to fulfill them.

Further, an AI decision-maker's conclusions are derived via probabilistic reasoning by drawing on large swathes of data that are not about any particular person. And typically the judgments they arrive at ultimately are not being made about a particular person either. I would be judged by my zip code, by my parents' actions, by the statistical likelihood that I would never recover, but as anyone with even a cursory understanding of statistics knows, probability tells us nothing about any individual case. While the use of big data to make decisions is certainly not new, its widescale use for life-changing decisions is. I deserve a judgment about my case, not a report about people like me in certain limited respects. I deserve to be seen as a particular individual, perhaps going through one of the worst experiences of my life. We do one another a disservice if in those conditions we choose to look away, to fail to acknowledge individual circumstances or recognize vulnerability, and instead treat one another as mere data points. Indeed, the very nature of such statistical judgments treats people as "kinds" deprived of individuality.

Nassim JafariNaimi (2018) makes a related point about the importance of attending to particular individuals when engaged in high-stakes decision-making, and about being mindful of the attitude toward those involved implied by the structure of decision-making processes themselves. In the quote below, JafariNaimi argues that using trolley-problem-style reasoning to arrive at generalizable moral principles for programming autonomous vehicles is deeply flawed on the grounds that it flattens and demeans living, breathing human beings

12. Though given both how they are designed and the degree to which humans have already anthropomorphized and formed attachments to AI chatbots (see, e.g, Andoh, 2025), we have good reasons for being concerned that we might be deceived into believing that they can.

13. Technically they say that "Robots cannot in principle possess these abilities" (852). However, they use 'robots' interchangeably with AWS in the paper.

with complex ends of their own. She writes in response to a proposal to drop a "fat man" into the path of an oncoming trolley in order to stop it and save other innocent victims:

> [W]e might first note how the fat man is treated as a living object, but an object nonetheless. All we know about him, after all, is his physical composition: he is fat and heavy, so much so that what is visible is the material weight of his body and the tragedy of his humanity disappears under this weight. The inhuman nature of this framing is clear when we try simple replacements: A fat girl? A heavy pregnant woman? My fat teenage brother? (313)

John Harris' (2020) scathing "The Immoral Machine" makes a similar point, though in Harris' case, he is responding to Edmond Awad et al.'s (2018) *Nature* article, "The Moral Machine Experiment." Harris argues that Awad et. al. trivialize the stakes of allowing autonomous vehicles to make life and death decisions when they suggest that we should crowdsource "public morality" to determine what life and death preferences are morally licensed by the public. Harris is primarily focused on the immorality of allowing generalizations about public preferences to license programming choices in AVs, evocatively making the point by noting:

> Having a preference for killing (or sparing) one person rather than another (even a preference shared with thousands or even millions) doesn't make it moral. The preferences of a certain 'Bohemian corporal,' it is salutatory to remember, came to be shared by millions. (Harris, 2020, p. 5)

Harris argues that treating life and death decisions like a series of "would you rather" trolley-problems completely fails to capture the stakes, and it eliminates the rich texture and detail that is part of any particular high-stakes decision.

While one might argue that JafariNaimi's (2018) and Harris's (2020) targets are more narrowly the misuse of trolley problems rather than a critique of high-stakes decision-making per se, the objections that they raise have as much to do with the attitudes expressed toward the individuals affected by the decisions as they have to do with the limitations of trolley problems. There are morally salient features to decision-making contexts that go beyond getting the right answer.

In sum, moral agency is a rich and multi-faceted capacity, one that AI does not possess, and showing recognition respect for the dignity of those about whom we make high-stakes decisions requires that they be made by a moral agent. This is even more true when there are identifiable agents (like doctors and judges) who actively stand in special relationships with those about whom the decisions are being made.

3.2 Effective Responsibility and Human Dignity

Up to this point in this section, I have primarily focused on why outsourcing moral judgments in high-stakes cases disrespects the human dignity of those about whom high-stakes decisions are being made. In the remainder of this section, I focus on how outsourcing our moral responsibilities, especially in high-stakes cases, degrades the dignity of the decision-makers. Given the role of citizens in collectively making high-stakes decisions for their shared societies, I suggest that normalizing more of this outsourcing runs the risk of decreasing our sense of effective responsibility for ourselves and for our societies by undermining our central human capacities and shared faith in the capacities of human beings (following Dewey, 1937/1981).

Even if AI judgments are the same that human judges would make given the laws as they are and given current precedent, needing to confront their application in particular cases keeps open the question about whether the laws and punishments in practice are the best ones. In his discussion about what might happen if the process of making and implementing laws were to become more automated and include less human judgment, Norman Spaulding offers the provocative claim that "in a free society there is no statement about what the law requires that is not also a question about whether it should be followed" (Spaulding, 2020, p. 400). His point is that the judgment about whether to obey a given law cannot be answered through the predictive methods of artificial intelligence. The decision to endorse a judgment and the decision to act are fundamentally the judgments of moral agents taking up responsibility in their own lives for the values that they hold dear, and often doing so in ways that increase their vulnerability.[14] These steps entail critical self-reflective attitudes, which AI currently lack, and we have good reason to doubt they will acquire in the near future. While Spaulding is specifically speaking to the legal context, I think this insight has far broader implications for robust moral responsibility and AI. Consistent human engagement in making these judgments and feeling the weight of them as, to an extent, on our own shoulders, encourages continued reflection about proportionality and about what punishments are compatible with human dignity. These reflective practices keep alive our engagement in the laws and norms of our societies.

As John Dewey (Dewey, 1937/1981) has characterized it, democracy requires faith in human capacities and in the potential for each person to contribute something of value to shared deliberation about the kind of society that we want to be. He claims that:

> [T]he key-note of democracy as a way of life may be expressed…as the necessity for the participation of every mature human being in formation of the values that regulate the living of men together—which is necessary from the standpoint of the general social welfare and the full development of human beings as individuals. (Dewey, 1937/1981, p. 218)

On Dewey's account, democracies start to break down when they become "too political in nature," when people cease having a sense of "effective responsibility" because they feel shut out and as though their input does not matter. While this article is not primarily about democracy, I take the attitude that Dewey expresses here to capture the respect and recognition that is at the core of treating one another as moral equals with dignity. Further, Dewey highlights how actively participating and collectively taking up responsibility for our societies is part of how we thrive as individuals. To cast off this responsibility and to see ourselves and others as having nothing of value to contribute is to treat ourselves as lesser than our dignity demands and to fail to cultivate or allow to degrade the practices and capacities needed for us to exercise robust moral agency.

While the circumstances of the exclusion that Dewey is describing are not primarily technological, this point analogizes well. If, rather than asking for human beings to participate, we instead have AI scrub our data to assess our interests, preferences, and desires, and then have it act on our behalf, it is not a stretch to anticipate that we might stop seeing ourselves as bearing this effective responsibility. It might become harder to see clear pathways for being able to shape our societies, yielding an increased sense of detachment from any of the decisions that are being made, and from our shared responsibility for the actions of our communities. Likewise, if we do not turn to others to learn from their experiences and insights, but we instead become increasingly more reliant on getting our information from

14. Consider, for example, someone choosing not to abide by the US Fugitive Slave Act, a group engaging in civil disobedience, or acts of jury nullification.

AI, this raises concerns for the good will toward one another that Dewey thinks is central to a flourishing democracy.

This flattening of human relationships is more concerning given climates of decreased social and institutional trust (see, e.g. Saad, 2023). Rather than continue to improve ourselves and commit to genuinely seeing one another, I fear that embracing AI as a way of responding to current lapses in equitable decision-making suggests giving up on the potential for us to do so and of our responsibility to try. Replacing more of our interpersonal interactions with machines seems unlikely to slow this trend, let alone rebuild some of the trust that has been lost.

One might respond that we do not need to be so disconnected from the judgments of AI. We can assign human beings to assess the output and to override it if human experts judge it to be mistaken, or if they think another interpretation of a case is appropriate. *Prima facie*, this seems like a reasonable compromise, a human being would be the one to endorse the judgment, to sign any relevant paperwork. But we should wonder if this is enough to protect the dignity of those affected by the decisions. Moreover, how likely is it that human judges would override the AI? In a 2019 Council of Europe report, Karen Yeung argues that:

> [I]ndividuals entrusted with the responsibility to supervise the operation of these systems may be understandably reluctant to intervene. This risks turning humans placed in the loop into 'moral crumple zones', largely totemic humans whose central role becomes soaking up fault, although they have only partial control of the system, and who are vulnerable to being scapegoated by tech developers and organisations seeking to avoid responsibility for unintended adverse consequences. (Yeung, 2019, p. 11)

Given both psychological and structural pressures to conform to the results, it seems unlikely that those granted the power to override would frequently do so, short of significant shifts in incentives that would likely work in tension with the efficiency benefits of AI. Likewise, as Yeung's provocative image suggests, even when people are held accountable for endorsing (or not) the AI output, this will likely fail to address the more fundamental issue of responsibility. This practice also still creates layers of distance between the decision-maker and those who will be directly affected by the decisions, generating a weaker sense of personal responsibility for some of the more challenging parts of shared social and political life.

For these reasons, I maintain that in at least some cases, the moral loss resulting from the use of AI in decision-making is sufficiently serious to warrant a right to a human decision. The corruption of human relationships and the degradation of human dignity may not be as extreme or acute in a one-off case, though I think it is a problem then too, but over time this more general practice is likely to warp how we see ourselves and one another. Where and how exactly we should draw these lines is a further question that I think we ought to answer through public deliberation and assessment of the values we want to protect and the risks we are morally willing to take in exchange for the advantages of AI. Engaging in public deliberation requires maintaining (and perhaps enhancing) the conditions of good deliberation and decision-making, and it is precisely these that automating more decisions to AI threatens to undermine.

4. Conclusion

Decisions about when and how to roll out sophisticated AI models are of course complex issues that require consideration of pragmatic challenges. In arguing that there is a moral loss when AI systems make important decisions about human lives, I do not mean that there are no cases when it might be appropriate for us to decide that, all things considered, this

is our best option. However, if we do so, we should be clear-eyed about the compromise we are making and aim to mitigate some of the worst consequences that might result from this practice. For instance, the use of AI to speed up the generation of policy is under way (Hisham et al., 2024). In some cases, the decision to use AI has been driven by the fact that it seemed better than the alternative of delayed policies and legal decisions. While that may be true in some cases, turning to AI as a solution, especially before it becomes more accurate and less biased, creates incentives to continue to do so, and disincentives to continue hiring and training humans to fill these roles.

Likewise, there might be cases where particular individuals would prefer that AI make the decision, for example, if they have themselves been subject to a pattern of biased treatment in a country with a history of inequitable application of the law. I do not mean to rule out the possibility that we might need to make concessions or that we should prevent people from having AI judges if that is what they want. However, the fact that a 2021 poll showed that over 50% of Europeans would support replacing some legislators with AI (IE, 2021) should not merely trigger an expansion of the use of AI, but also some serious internal reflection about how we have arrived at this point, and to what extent greater divisions between people and mistrust have been part of that process. If I am right that these are significant contributing factors to current social problems, we should be hesitant to deploy AI in the most vulnerable parts of our shared lives rather than invest more energy into resolving some of these underlying challenges.

In conclusion, advancements in AI open incredible possibilities. However, if deployed to make high-stakes decisions about the lives of individuals with whom we stand in significant relationships, we outsource a crucial human responsibility. In doing so, I have argued, we undermine both the dignity of the person about whom the decision is made and that of ourselves when we fail to take up our responsibilities. I have argued that these problems remain even if we make significant progress on issues of bias, transparency, and legal accountability. Determining which contexts are so morally important that they warrant protection from AI judgments requires further public discussion. However, they are likely those that concern our greatest human vulnerabilities, those about our liberty, our health, and our relationships. We well might decide that given our very imperfect world, there are some scenarios in which these tradeoffs are the best option. Nonetheless, I argue that we must keep front-and-center that moral agency requires us making choices, acting upon them, and bearing their weight.

Bibliography

Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey. *IEEE Access*, *13*, 18912–18936. https://doi.org/10.1109/ACCESS.2025.3532853

Andoh, E. (2025). Many teens are turning to AI chatbots for friendship and emotional support. *Https://Www.Apa.Org*, *56*(7). https://www.apa.org/monitor/2025/10/technology-youth-friendships

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. *Science*, *373*(6552), 284–286. https://doi.org/10.1126/science.abg1834

Buslón, N., Cortés, A., Catuara-Solarz, S., Cirillo, D., & Rementeria, M. J. (2023). Raising awareness of sex and gender bias in artificial intelligence and health. *Frontiers in Global Women's Health*, *4*, 970312. https://doi.org/10.3389/fgwh.2023.970312

Crawford, K. (2021). *Atlas of AI: power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Darwall, S. L. (1977). Two Kinds of Respect. *Ethics*, *88*(1), 36–49.

Denworth, L. (2019, October 1). *The Significant Problem of P Values*. Scientific American. https://www.scientificamerican.com/article/the-significant-problem-of-p-values/

Deroy, O. (2023). The Ethics of Terminology: Can We Use Human Terms to Describe AI? *Topoi*, *42*(3), 881–889. https://doi.org/10.1007/s11245-023-09934-1

Dewey, J. (1981). Democracy and Educational Administration. In J. A. Boydston (Ed.), *John Dewey: The Later Works, 1925-1953* (Vol. 11, pp. 217–225). Southern Illinois University Press : Feffer & Simons. (1937)

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Equivant Supervision. (2024, August 12). Debunking Misconceptions About the COMPAS Core Instrument: What You Need to Know. *Equivant Supervision*. https://equivant-supervision.com/debunking-misconceptions-about-the-compas-core-instrument-what-you-need-to-know/

Garcia, A. C. B., Garcia, M. G. P., & Rigobon, R. (2024). Algorithmic discrimination in the credit domain: what do we know about it? *AI & SOCIETY*, *39*(4), 2059–2098. https://doi.org/10.1007/s00146-023-01676-3

Gow, G. (2022, July 17). *How To Use AI To Eliminate Bias*. Forbes. https://www.forbes.com/sites/glenngow/2022/07/17/how-to-use-ai-to-eliminate-bias/?sh=7e3c3bd91f1f

Gunkel, D. J. (2020). Mind the gap: responsible robotics and the problem of responsibility. *Ethics and Information Technology*, *22*(4), 307–320. https://doi.org/10.1007/s10676-017-9428-2

Harris, J. (2020). The Immoral Machine. *Cambridge Quarterly of Healthcare Ethics*, *29*(1), 71–79. https://doi.org/10.1017/S096318011900080X

Hattingh, J. (2025). New AI Transparency Rules Have a Trade Secrets Problem. *Lawfare*. https://www.lawfaremedia.org/article/new-ai-transparency-rules-have-a-trade-secrets-problem

Heaven, W. D. (2022). Bias Isn't the Only Problem with Credit Scores—and No, AI Can't Help *. In *Ethics of Data and Analytics*. Auerbach Publications.

Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis. *Science and Engineering Ethics*, *21*(3), 619–630. https://doi.org/10.1007/s11948-014-9565-5

Hindriks, F., & Veluwenkamp, H. (2023). The risks of autonomous machines: from responsibility gaps to control gaps. *Synthese*, *201*(1), 21. https://doi.org/10.1007/s11229-022-04001-5

Hisham, A. 'Aisha B., Yusof, N. A. M., Salleh, S. H., & Abas, H. (2024). Transforming Governance: A Systematic Review of AI Applications in Policymaking. *Journal of Science, Technology and Innovation Policy*, *10*(1), 7–15. https://doi.org/10.11113/jostip.v10n1.148

Huq, A. (2020). *A Right to a Human Decision*. *106*(3). https://virginialawreview.org/articles/right-human-decision/

IE. (2021, May 27). *IE University Research on Replacing MPs with Robots Reveals 1 in 2 Europeans Want to Replace National MPs with Robots.* IE University. https://www.ie.edu/university/news-events/news/ie-university-research-reveals-1-2-europeans-want-replace-national-mps-robots/

JafariNaimi, N. (2018). Our Bodies in the Trolley's Path, or Why Self-driving Cars Must *Not* Be Programmed to Kill. *Science, Technology, & Human Values*, *43*(2), 302–323. https://doi.org/10.1177/0162243917718942

Jasanoff, S. (2016). *The ethics of invention: technology and the human future* (First edition). W.W. Norton & Company.

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, *160*, 108352. https://doi.org/10.1016/j.chb.2024.108352

Kostick-Quenet, K. M., Cohen, I. G., Gerke, S., Lo, B., Antaki, J., Movahedi, F., Njah, H., Schoen, L., Estep, J. E., & Blumenthal-Barby, J. S. (2022). Mitigating Racial Bias in Machine Learning. *Journal of Law, Medicine & Ethics*, *50*(1), 92–100. https://doi.org/10.1017/jme.2022.13

Laker, B. (2023, July 7). *The Dark Side Of AI Recruiting: Depersonalization And Ghosting*. Forbes. https://www.forbes.com/sites/benjaminlaker/2023/07/07/the-dark-side-of-ai-recruiting-depersonalization-and-its-consequences-on-the-modern-job-market/?sh=31db580e68c8

Lamberton, C., Ruster, L., Ghai, S., & Saldanha, N. (2025). Dignity, properly used, could be a useful construct in AI ethics. *Patterns*, *6*(11). https://doi.org/10.1016/j.patter.2025.101396

Lee, P. (2016, March 25). Learning from Tay's introduction. *The Official Microsoft Blog*. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

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

May, J. R., & Daly, E. (2019). Why dignity rights matter. *EUROPEAN HUMAN RIGHTS LAW REVIEW*, (2), 129–134. https://doi.org/10.3316/agispt.20190424009530

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

Metz, C., & Weise, K. (2025, May 5). A.I. Is Getting More Powerful, but Its Hallucinations Are Getting Worse. *The New York Times*. https://www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html

Moschella, D. (2022). *AI Bias Is Correctable. Human Bias? Not So Much*. https://itif.org/publications/2022/04/25/ai-bias-correctable-human-bias-not-so-much/

Nussbaum, M. C. (2007). *Frontiers of justice: disability, nationality, species membership* (1. Harvard Univ. Press paperback ed). Belknap Press.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

O'Brien, M. (2024, February 23). *Google says AI image-generator would sometimes "overcompensate" for diversity | AP News*. AP News. https://apnews.com/article/google-gemini-ai-chatbot-imagegenerator-race-c7e14de837aa65dd84f6e7ed6cfc4f4b

O'Neil, C. (2016). *Weapons of math destruction: how big data increases inequality and threatens democracy* (First edition). Crown.

Parliamentary Council. (1949). *Basic Law for the Federal Republic of Germany*. https://www.gesetze-im-internet.de/englisch_gg/englisch_gg.html

Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous Machines, Moral Judgment, and Acting for the Right Reasons. *Ethical Theory and Moral Practice*, *18*(4), 851–872. https://doi.org/10.1007/s10677-015-9563-y

Raghavan. (2024, February 23). *Gemini image generation got it wrong. We'll do better.* Google. https://blog.google/products-and-platforms/products/gemini/gemini-image-generation-issue/

Saad, L. (2023, July 6). *Historically Low Faith in U.S. Institutions Continues*. Gallup.Com. https://news.gallup.com/poll/508169/historically-low-faith-institutions-continues.aspx

Sandel, M. J. (2012). *What money can't buy: the moral limits of markets*. Farrar, Straus and Giroux.

Silberg, J., & Manyika, J. (2019, June 6). *Tackling Bias in Artificial Intelligence (And Humans)*. McKinsey Global Institute.

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Sparrow, R. (2021). Why machines cannot be moral. *AI & SOCIETY*, *36*(3), 685–693. https://doi.org/10.1007/s00146-020-01132-6

Spaulding, N. (2020). Is Human Judgment Necessary?: Artificial Intelligence, Algorithmic Governance, and the Law. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford handbook of ethics of AI* (pp. 375–401). Oxford University Press.

Tasioulas, J. (2022). Artificial Intelligence, Humanistic Ethics. *Daedalus*, *151*(2), 232–243. https://doi.org/10.1162/daed_a_01912

UN General Assembly. (1948). *Universal Declaration of Human Rights*. United Nations. United Nations. https://www.un.org/en/about-us/universal-declaration-of-human-rights

Venzke, O., & Akselrod, C. (2023, August 23). How Artificial Intelligence Might Prevent You From Getting Hired | ACLU. *American Civil Liberties Union*. https://www.aclu.org/news/racial-justice/how-artificial-intelligence-might-prevent-you-from-getting-hired

Vicente, L., & Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific Reports*, *13*(1), 15737. https://doi.org/10.1038/s41598-023-42384-8

Victoroff. (2022, February 5). *Turtles all the way down: Why AI's cult of objectivity is dangerous, and how we can be better*. Venturebeat. https://venturebeat.com/ai/turtles-all-the-way-down-why-ais-cult-of-objectivity-is-dangerous-and-how-we-can-be-better

Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., Zhao, D., Shirai, I., Narayanan, A., & Russakovsky, O. (2022). REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. *International Journal of Computer Vision*, *130*(7), 1790–1810. https://doi.org/10.1007/s11263-022-01625-5

Yeung, K. (2019). *Responsibility and AI: Council of Europe Study DGI(2019)05*. https://edoc.coe.int/en/artificial-intelligence/8026-responsibility-and-ai.html