JOHN DANAHER

IDEAS FEB 2, 2023 9:00 AM

The Case for Outsourcing Morality to AI

As AI infiltrates more aspects of society, maybe some “responsibility gaps” are a good thing.

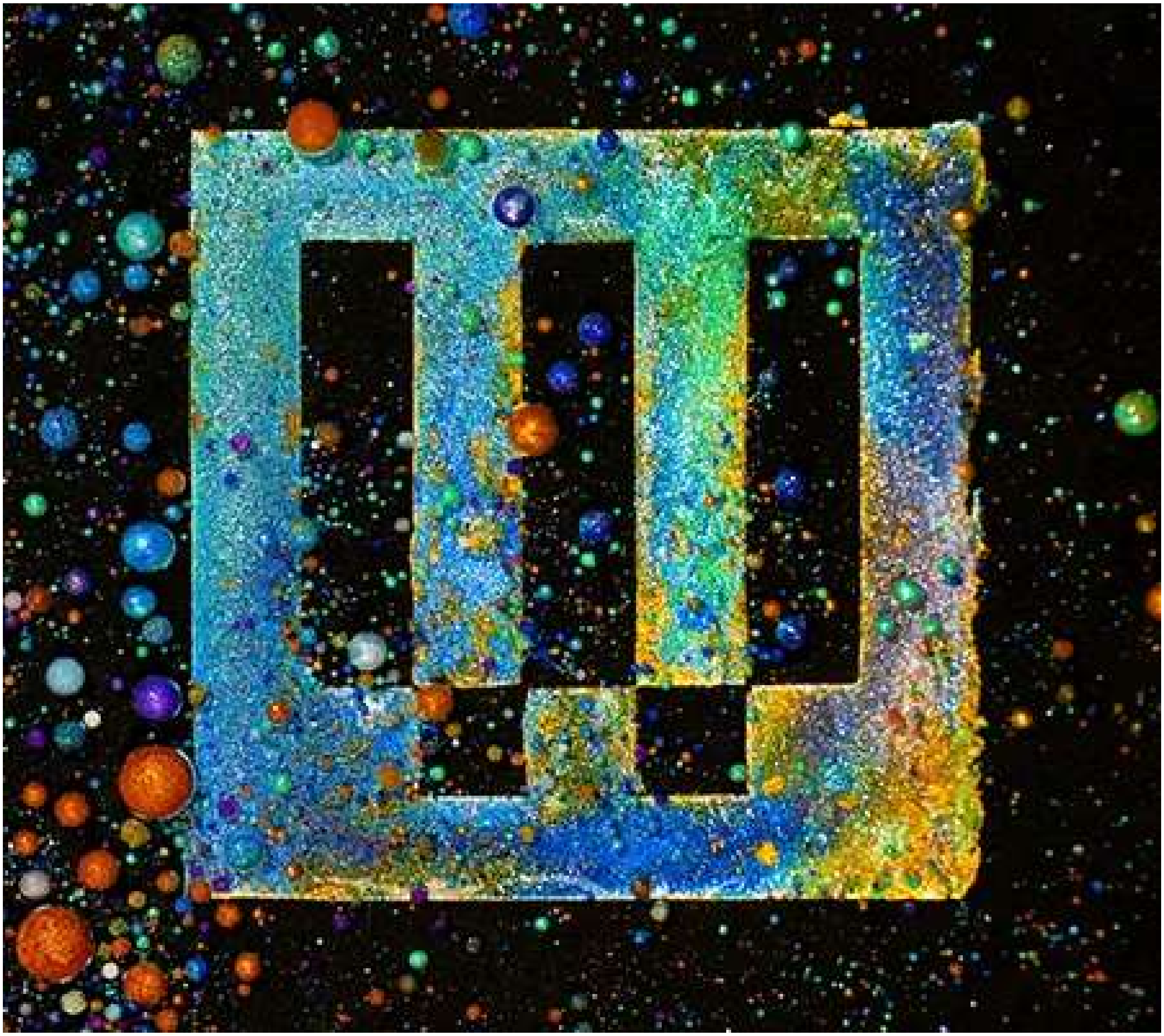
PHOTO-ILLUSTRATION: WIRED STAFF; GETTY IMAGES

SAVE

IT ALL STARTED with an obscure article in an obscure journal, published just as the last AI winter was beginning to thaw. In 2004, Andreas Matthias wrote an article with the enigmatic title, “The responsibility gap: Ascribing responsibility for the actions of learning automata.” In it, he highlighted a new problem with modern AI systems based on machine learning principles.

Once, it made sense to hold the manufacturer or operator of a machine responsible if the machine caused harm, but with the advent of machines that could learn from their interactions with the world, this practice made less sense. Learning automata (to use Matthias' terminology) could do things that were neither predictable nor reasonably foreseeable by their human overseers. What's more, they could do these things without direct human supervision or control. It would no longer be morally fair or legally just to hold humans responsible for the actions of machines. Matthias argued that this left humanity in a dilemma: Prevent the development of learning automata or embrace the responsibility “gaps” that resulted from their deployment.

Fast forward to 2023 and Matthias' dilemma is no longer of mere academic concern. It is a real, practical issue. AI systems have been, at least causally, responsible for numerous harms, including discrimination in AI-based sentencing and hiring, and fatal crashes in self-driving vehicles. The academic and policy literature on "responsibility gaps" has unsurprisingly ballooned. Matthias' article has been cited over 650 times (an exceptionally high figure for a philosophy paper), and lawyers and policymakers have been hard at work trying to clarify and close the gap that Matthias identified.



What is interesting about the responsibility gap debate, however, is the assumption most of its participants share: that human responsibility is a good thing. It is a good thing that people take responsibility for their actions and that they are held responsible when something goes wrong. Contrariwise, it would be a bad thing if AI systems wreaked havoc in the world without anyone taking responsibility or being held responsible for that havoc. We must, therefore, find some way to plug or dissolve responsibility gaps, either by stretching existing legal/moral standards for responsibility, or introducing stricter standards of responsibility for the deployment of AI systems.

But perhaps responsibility is not always a good thing. Perhaps, to follow Matthias's original suggestion, some responsibility gaps ought to be embraced.

It is worth bearing in mind two features of our world. First, our responsibility practices (as in, our norms and habits of blaming, shaming, and punishing one another) have their dark side. Second, our everyday lives are replete with "tragic choices," or situations in which we have to choose between two morally equal or close-to-equally-weighted actions. Both features have implications for the responsibility gap debate.

On the dark side of responsibility, an entire school of thought has emerged that is critical of our responsibility practices, particularly as they pertain to criminal justice. Gregg Caruso, a philosophy professor at the State University of New York, is one of the leading lights in this school of thought. In conversation with me, he argued that if you "look closely … you will find that there are lifetimes of trauma, poverty, and social disadvantage that fill the prison system." Unfortunately, our current responsibility practices, premised on the ideal of free will and retributive justice, does nothing to seriously address this trauma. As Caruso put it, this system "sees criminal behavior as primarily a matter of individual responsibility and ends the investigation at precisely the point it should begin." If we abandoned our system of retributive justice, we could "adopt more humane and effective practices and policies." Caruso also pointed that our emotions associated with responsibility—what philosophers call 'reactive attitudes' such as resentment, anger, indignation, and blame, are "often counterproductive and corrosive to our interpersonal relationships" because they "give rise to defensive or offensive reactions rather than reform and reconciliation."
Of course, defenders of our responsibility practices could respond by claiming that as long as we correctly identify the guilty, and fairly apportion blame, all the suffering and trauma that Caruso highlights is besides the point. Punishment is supposed to be harsh and, in a sense, traumatic. This, however, ignores the growing evidence to suggest that we are often too willing to blame people, even when the facts may not justify our desire to do so. Studies by the psychologist Mark Alicke, for instance, suggest that people often engage in *blame validation*, meaning that first they find someone to blame, then they find a way to justify it. Collectively, this evidence, when tied to Caruso's arguments, suggests that our current responsibility practices can be morally inhumane and cause unnecessary scapegoating, physical harm, and psychological torment.

Additionally, a number of philosophers have highlighted the tragic nature of our moral choices. Lisa Tessman, from Binghamton University, is one of the most articulate and emphatic defenders of the idea. In her books, *Moral Failure* and *When Doing the Right Thing is Impossible*, she highlights numerous moral dilemmas and choices we face in life, each of which involves some unavoidable and hard-to-evaluate tradeoff between competing moral considerations. Here's a simple example: Imagine that you are a parent to two children. You love them both and think they are both equally morally deserving of your attention and love. Nevertheless, the world being the way it is, you will frequently have to pick and choose between them, attending one child's soccer match while missing the other's piano recital (or some variation on this theme). This is what it means to face a tragic choice: to be forced to pick between incommensurable and/or equally valid moral considerations. How common is this phenomenon? As Tessman put it to me, moral intuition often leads us "to the verdict that we are impossibly required to do something, such as protect a loved one, even if we are unable to do so, or carry out both of two, non-negotiable moral requirements." So common is the experience, in fact, that Tessman takes "human life to be full of tragedy" because "humans are vulnerable to losing what we deeply value and cannot replace … [and] we are often in situations in which we cannot protect others from these losses."

The parent-child example is a relatively low stakes and private instance of tragic choice. There are many high-stakes, public decisions that involve similar tradeoffs. Consider decisions about the allocation of scarce medical resources (the "Who gets the ventilator?" dilemma that arose early in the Covid-19 pandemic) or the allocation of social opportunities (scholarships, funding). Anyone who has been involved in such decisions will know that they often devolve into largely arbitrary choices between equally deserving candidates. While some people can ignore the apparent tragedy inherent in such decisions, others anguish over them. Tessman argues that this anguish is a "fitting" response to the pervasiveness of tragedy. But some responses are not so fitting: To morally blame people for their choices in such contexts, and to punish them for making what you think is the wrong choice, is perverse and unjustified. And yet people often cannot resist the urge to do so.

These two considerations—that responsibility has a dark side and tragic choices are commonplace—give us reason to embrace at least some responsibility gaps. To be more precise, in any decisionmaking context in which a) we face a tragic choice; b) holding a human responsible in such a context would risk unnecessary scapegoating; and c) the AI system would be capable of making the same kinds of decision as a human decisionmaker, we have good reasons to favor delegation to machines, even if this means that nobody can be held responsible for the resulting outcomes.

To put the case another way: Holding one another responsible has psychological and social costs. In at least some cases, imposing those costs is morally unjustified. If, in those cases, we can delegate decisions to machines, and those machines are not obviously "worse" than humans at making those decisions, then why shouldn't we do so?

Objections to this proposal are likely to come thick and fast.

First, some people might argue that the proposal is not psychologically possible. People won't buy it. The instinct to find a human scapegoat is too strong. But there is some initial empirical evidence to suggest that people would be receptive. Matthias Uhl, a behavioral psychologist at the Technical University of Ingolstadt in Germany, has been studying this phenomenon for some time. He says that ordinary people have “no problem” with assigning responsibility to AI systems, even if “ethicists consider this a category mistake.” Furthermore, in a recent study with his colleagues, he found that people could reduce their responsibility by delegating decisions to machines. The study allowed people to “transfer a task with potentially detrimental consequences for another person … to a machine or do it themselves.” Uhl and his colleagues found that “if the machine failed … [the human delegators] were punished less severely by the person that was harmed than if they failed themselves.” The same effect did not arise if they delegated the task to another human. Uhl, who is quick to point out that this finding needs to be more robustly validated, nevertheless suggests that the experiment does “seem to be evidence that people might be able to successfully reduce perceived moral responsibility by delegation to machines.”

Others might object that the proposal is not morally possible. Even if people are less willing to punish others in the case of delegation to machines, they are not morally justified in doing so. If I choose to delegate some tragic choice to an AI—such as the choice to allocate scarce medical equipment—I am still responsible for making that choice because I made the choice to delegate. People can rightly blame me for that and for the consequences of delegation. Moral responsibility isn't eliminated; it is just pushed one step further back.

There are, however, two problems with this objection. First, even if there is responsibility for the decision to delegate, it is of a different character to the responsibility to allocate the medical equipment. The delegator cannot be blamed for the particular allocation that the AI system comes up with. There is a net reduction in the overall level of blame and the unnecessary suffering and punishment that could result from the decision to allocate. Morally justifiable blame is reduced, if not eliminated. Second, the objection misses the point. The whole point of the argument is that there are some cases in which it is unfair and morally costly to put humans “on the hook” for the decision. Delegation should be an option in those cases.
Finally, some might object that welcoming responsibility gaps in this instance would be to take the first step down a slippery slope. What about devious actors that want to avoid responsibility for their actions? As some have put it, there is a real risk of corporate and governmental actors “laundering” their moral and legal responsibility through machines. Indeed, we see this happening already. Consider Facebook’s disavowals of responsibility for malicious or hateful content that people see on their platform. When challenged about this, they will try to correct the problem but will argue that it’s not them, but the algorithm. But aren’t they the ones that choose to create and monetize a platform with a particular algorithmic regulation? Aren’t they rightly held responsible for this? And doesn’t Uhl’s experiment suggest that there is a real danger that people will not do so?

This, to me, is the most serious objection to the proposal. Even if there are some cases in which responsibility gaps are welcome, there are also some (perhaps many) cases in which they are not. Getting the balance right will be tricky but the risk of unwarranted delegation is not a reason to reject warranted delegation. I would not advocate reckless or universal outsourcing to machines; we should be thoughtful and careful in choosing the right contexts in which to outsource. Periodic review and auditing of our decision to outsource would be appropriate. Still, this is not a reason to never outsource. Nor is it a reason to close all responsibility gaps. The tragic nature of many moral choices and the zeal to overpunish still give us reasons to embrace them in some cases.

- The AI detection arms race is on
- The gruesome story of how Neuralink’s monkeys actually died
- Quan Millz was the biggest mystery on TikTok—until now
- There’s an alternative to the infinite scroll
- AI chatbots are invading your local government
- 🔌 Charge right into summer with the best travel adapters, power banks, and USB hubs

John Danaher is Senior Lecturer at the School of Law, University of Galway, Ireland. He is the author of Automation and Utopia (Harvard University Press 2019) and the co-author of A Citizen's Guide to Artificial Intelligence (MIT Press, 2021). He is currently writing a book on why and how to... Read more

TOPICS ETHICS PHILOSOPHY ARTIFICIAL INTELLIGENCE

MORE FROM WIRED

Preferring Biological Children Is Immoral

Most people say they want their kids to be their own genetic offspring—but such a desire is in conflict with other evolving values around parenting and family.

LEO KIM

The Hollywood Writers AI Deal Sure Puts a Lot of Trust in Studios to Do the Right Thing

The Writers Guild of America won important protections, but it's not enough. When the Screen Actors Guild goes to the table, it should fight for more to keep AI from impinging on the work of artists.

ALEX WINTER

Immersive Tech Obscures Reality. AI Will Threaten It

AI could supercharge augmented and virtual reality, making online manipulation and disinformation campaigns much more personal—and effective.

JAMESON SPIVACK

AI-Powered ‘Thought Decoders’ Won’t Just Read Your Mind—They’ll Change It

“Mind-reading” neural decoders could spell the end of privacy. But the full ramifications of this technology are even more concerning.

LEO KIM

The Hypocrisy of Judging Those Who Become More Beautiful

Normally, correcting disadvantages beyond our control is seen as laudable. So why do people look down on individuals who alter their looks?

SHEON HAN

Everyone Is a Girl Online

NPC influencers, "girl dinner," angels, bimbos—the internet is a girl's world now, whether you like it or not.

ALEX QUICHO

I Failed Two Captcha Tests This Week. Am I Still Human?

WIRED’s spiritual advice columnist on whether modern tech makes people behave more like bots.

MEGHAN O'GIEBLYN

How a ‘Digital Peeping Tom’ Unmasked Porn Actors

“It’s like the secret identity of Batman or Superman. You’re not supposed to know who this person is, they didn’t want you to know, and somehow you found out.”

KASHMIR HILL