AI ETHICS - ESSAY

We Built the Monster We're Afraid Of

Why our fears of AI say more about us than they do about the technology.

Mandy Hathaway - MA, Ethical Technology & Artificial Intelligence, Metropolitan State University

The Paradox

Here is something worth sitting with: we trust AI more when it seems human, and then we fear it more for exactly the same reason.

We give our digital assistants human names -- Alexa, Siri, Watson. We train them to speak in natural sentences, to pause at the right moments, to apologize when they get something wrong. We do this because it works. Research consistently shows that when people perceive a system as having a social presence, their comfort and engagement with it increase. The more human it seems, the more we trust it.

And then we go home and watch movies about the robots taking over.

This is not a coincidence. The same psychological mechanism that makes us warm to AI -- our deep, evolved tendency to project human qualities onto the things around us -- is also the engine behind our most dramatic fears about it. We look at the systems we have built, see ourselves reflected back, and then wonder what they might do with that power. The fear is real. But it is a fear of us, not of them.

Why We See Ourselves Everywhere

Anthropomorphism -- projecting human qualities onto non-human things -- is one of the most fundamental features of human psychology. It predates written language. It is in our mythology, our religion, our children's books, and our relationship with our cars. We name boats. We apologize to furniture we walk into. We feel genuinely guilty about throwing away a childhood toy.

With animals, this tendency has some scientific backing. Biological brains, however different in size and structure, share a common substrate: they are all subject to hormones, neurotransmitters, and the chemical cocktails that produce fear, pleasure, attachment, and grief. When Carl Safina argues that if we can assume animals are tired when they sleep and hungry when they eat, we can reasonably assume they are happy when they play -- that is not sentimentality. That is a reasonable inference from shared biology.

Machines are different. A thermostat does not suffer when the house is cold. A car does not need a rest after a long drive. When we project emotional states onto objects, we are not picking up on signals the object is sending -- we are filling a silence with our own interior life. The object is a screen, and what we see on it is ourselves.

> *When we anthropomorphize machines, we are not reading them. We are writing ourselves onto them. The question is what we choose to write.*

AI Is Designed to Trigger This

What makes AI different from the furniture we apologize to is that AI is deliberately engineered to invite anthropomorphism. This is not accidental and it is not subtle. Systems are trained on vast databases of human-generated language specifically so they can mimic the patterns of human communication. They are given voices calibrated to sound warm. They are assigned names that feel familiar. The entire design philosophy is oriented around one goal: make the user feel like they are interacting with something that understands them.

It works extraordinarily well. Studies show that people working alongside robots attributed more responsibility to the robot for shared work when they had been primed to think of it as human-like. People are less likely to replace a device they have anthropomorphized -- not because the device is performing better, but because replacing it feels like abandoning a friend. Brand loyalty, in the age of AI, is partly a psychological artifact of good interaction design.

Companies are aware of this and lean into it. 'Attributing a mind to a machine matters,' as one research team put it, 'because it could create a machine to which users might entrust their lives.' That is both the appeal and the risk, delivered in the same sentence.

The same mechanism that makes people comfortable enough to use these systems also makes them willing to trust those systems beyond what the technology warrants. We explored this problem in depth in the companion essay on algorithmic bias -- judges trusting COMPAS scores, drivers ignoring Tesla's explicit warnings to stay alert. Anthropomorphism is not just a psychological curiosity. It is an active ingredient in how these systems cause harm.

The Terminator Problem

From Stephen Hawking to Elon Musk, serious people have warned that advanced AI could spell the end of the human race. This has spawned think tanks, academic centers, and a flood of cultural product -- The Terminator, The Matrix, I, Robot, Black Mirror, and dozens more -- all exploring some version of the same scenario: machines become intelligent, machines turn on us, machines win.

These scenarios generally hinge on what researchers call the AI Singularity: the idea that once AI reaches human-level intelligence, it will be capable of improving itself, producing systems smarter than itself, producing systems smarter than those, and so on -- accelerating beyond any possibility of human control. At that point, the story goes, all bets are off.

Let us take these scenarios seriously and look at what they actually require.

The accidental apocalypse -- the paperclip scenario, where a superintelligent system tasked with making paperclips turns everything in the world into paperclips, including the humans -- does not actually implicate the AI at all. It implicates the designers. A system powerful enough to reshape the world does not spring into existence overnight. It begins the way all AI begins: as a narrow system, built by human programmers, trained on human-curated data, deployed with human-defined objectives. Every step of that process involves human decisions. The catastrophic outcome in Bostrom's scenario is a failure of human foresight, not evidence of machine malevolence.

The rebellion scenario -- machines developing consciousness, being denied rights, and eventually fighting back -- places the moral responsibility squarely on humans as well. In that story, the AI is not the aggressor. It is a subjugated population defending itself. That is a story about what humans do to things they have power over, not a story about what AI does when left unchecked.

Then there is the Frankenstein scenario: the machine simply turns on its creators. This one requires the most examination, because it is the most psychologically revealing.

Whose Psychology Is This, Really?

Harvard psychologist Steven Pinker has a theory about where Frankenstein fears come from. He argues they result from confusing high intelligence with megalomaniacal goals -- what he describes as 'a projection of alpha-male psychology onto the very concept of intelligence.'

That framing is worth sitting with. The assumption embedded in almost every AI apocalypse scenario is that a sufficiently intelligent system will want power, will want dominance, will want to control or destroy the things around it. But where does that assumption come from? The desire to dominate, to subjugate, to crush competitors -- these are not properties of intelligence. They are properties of a particular kind of evolved biological organism under particular kinds of social and survival pressures.

The desire for power did not emerge in humans because we are smart. It emerged because our ancestors lived in environments where social dominance improved survival odds. It is a biological inheritance, rooted in hormone systems and neural architecture shaped over millions of years. It is, in the most literal sense, our animal nature -- and it is something that human reason and social development work constantly to manage and restrain.

A non-biological superintelligence would have none of that history. No cortisol spikes. No testosterone-driven status competition. No fear of death to motivate self-preservation at others' expense. The drives that make power-hungry humans dangerous are chemical and evolutionary in origin. There is no logical reason to assume they would be emergent properties of a digital system, no matter how intelligent.

> *We assume that if a machine becomes smart enough, it will want what the worst of us want. That assumption tells us something true -- just not about the machines.*

And there is one more piece worth naming. Studies consistently find that psychopathic traits -- narcissism, aggression, a desire to dominate, an absence of empathy -- are dramatically overrepresented in corporate leadership. While people displaying these traits make up roughly 1% of the general population, research suggests they may account for up to 20% of C-suite executives. In societies that reward these behaviors and normalize them as strength, it is not surprising that we project them onto the most powerful things we create. We built the monster in our own image. The image we chose was not our best one.

The Risks That Actually Deserve Our Attention

None of this is an argument for complacency about AI development. Quite the opposite.

The danger of fixating on apocalyptic scenarios is that it pulls attention away from the harms that are happening right now, to real people, in systems already deployed. Biased risk assessment algorithms sending people to prison for longer than they deserve. Hiring systems filtering out qualified candidates based on demographic proxies embedded in training data. Recommendation engines amplifying misinformation at scale. Facial recognition systems with dramatically higher error rates for dark-skinned faces, used by law enforcement agencies that trust the outputs anyway.

These harms do not require a Singularity. They do not require consciousness or rebellion or any of the narrative apparatus of science fiction. They require nothing more than what already exists: systems trained on biased data, deployed without adequate oversight, by organizations that profit from the deployment and bear little cost when things go wrong.

Technological progress works like biological evolution in one important respect: it is iterative. Nothing springs fully formed from a designer's imagination. Today's AI systems are built on yesterday's, and tomorrow's will be built on today's. The biases and blind spots we allow to persist in current systems do not disappear when we build the next generation -- they become the foundation. Every problem we fail to address now gets inherited.

The future of AI will not be determined by whether we can prevent a machine uprising. It will be determined by the choices we make right now about what we build, who we include in building it, what we optimize for, and who we hold accountable when it causes harm. Those are human questions. They have human answers. And they are already overdue.

--

Mandy Hathaway *holds an MA in Ethical Technology & Artificial Intelligence and a BA in Philosophy from Metropolitan State University. She works at the intersection of AI systems, technical communication, and ethics.*