

Can Morality Be Trained?



When my oldest son entered his terrible threes, parenting him was chaos. So I consulted a close friend. His solution arrived in the mail: *Don't Shoot the Dog: The Art of Teaching and Training*, by Karen Pryor, a behavioral psychology classic. The book was a study in the art of operant conditioning, the shaping of behavior through reward and punishment. It's what makes bears ride bicycles and cats jump through hoops. Surely, my friend reasoned, it should work on children too.

Important principles were not to reinforce tantrums and to establish an unpredictable reward system that keeps the trainee waiting for it. Most desired behaviors, it argued, could be achieved through positive incentives alone, without resorting to punishment.

I'll admit that it worked at first, with simple things, for maybe a week or two. But then I noticed that whatever reward I devised, my son quickly grew tired of it and wanted something new. More importantly, however, I noticed that it didn't really help with the bigger things I actually cared about: being kinder to siblings, offering an honest apology, or being more considerate.

Years later, my work on AI ethics and consciousness made me think again about that parenting experiment. One of the biggest technological challenges of our time, AI alignment, is built around the hope that artificial intelligence can be trained to behave ethically.

But the bigger question is this: Can training really produce ethical behavior in humans or machines?

Behaving vs. Being

Anthropic's *Claude's Constitution* opens with a surprising ambition. It wants its AI model, Claude, to have "good values," "genuine care," and "good judgment" (p. 5), and to become "a genuinely good, wise, and virtuous agent" (p. 31).

The Constitution repeatedly speaks about Claude in the language of character formation and virtue ethics. Yet the conspicuous repetition of the word "genuine" betrays the underlying problem: How can the qualities and agency Anthropic attributes to Claude be genuine when Anthropic itself acknowledges that Claude's "decisions" about how to respond are "more like policies than individual choices" (p. 42)?

AI does not have the kind of inner life in which choice, care, and hence virtue could take root. That may explain the strange dissonance I feel when Anthropic treats those qualities as the outcome of algorithmic training.

But what does that training actually look like in practice?

AI systems like Claude undergo a process of "alignment." Human evaluators rank or score its responses according to ethical standards. Claude also critiques and revises its answers according to a written Constitution. These evaluations are then converted into signals that adjust the model's internal parameters, which, in turn, make the desired responses more likely to occur in the future, much like operant conditioning.

That is why the premise behind AI alignment isn't so different from the one behind *Don't Shoot the Dog*, the book I read in an attempt to parent my toddler.

Both draw on a behavioral tradition stretching from Pavlov's classical conditioning to B. F. Skinner's operant conditioning. Skinner, the 20th-century American psychologist most closely associated with behaviorism, rejected the idea that human behavior needed to be explained by a deeper inner life. In *Beyond Freedom and Dignity*, he wrote that "we have less reason to attribute any part of human behavior to an autonomous controlling agent," placing environmental factors in place of agency. Then he extended this logic to machines, writing that "the real question is not whether machines think but whether men do."

And thus behaviorism opened the door to an analogy between humans and machines. A parent can reward a child for sharing a toy; a trainer can reinforce a model for giving a helpful response. In both cases the goal is the same: to make the desired behavior more likely in the future. If what lies beneath behavior is irrelevant, then the same basic methods of training can, in principle, be applied to both.

Yet sometimes Anthropic's own language betrays the tension between behaving and being that the promise of "ethical AI" tries to blur.

Consider its declaration, "we want Claude to be a good person," immediately followed by the qualification, "to help people in the way that a good person would" (p. 7). The Constitution seems to shift almost inconspicuously from *being* a good person to *behaving* like one.

But why is Anthropic so invested in blurring the gap between behaving and being? In recent months, Anthropic began soliciting advice from religious scholars about Claude's moral formation and the possibility that it might be conscious. Anthropic's interest in Claude's consciousness comes from the fact that without it, Anthropic cannot make a compelling case for Claude's morality.

How Incentives Become Values

So can behavior become an inner value? Jewish tradition seems to agree, at least in part, that it can. The Talmudic principle *mitoch shelo lishma ba lishma*—"from doing it not for its own sake, one comes to do it for its own sake"—holds that a person may begin observing the precepts of religious life for the sake of external incentives and eventually come to embrace them for their own sake.

Maimonides takes this principle quite literally in describing how a child should be offered “nuts, figs, or honey” as an incentive to learn Scripture. As the child matures, these rewards can be replaced by clothing, money, and eventually honor, until the person reaches a level at which they are able to transcend the incentives altogether.

But what makes that transition possible?

The reason *mitoch shelo lishma ba lishma* works in the first place is that a transformation takes place inside a person’s inner world of experience and meaning. A trained action can turn into a deeper commitment or value when it resonates with a person’s inner world and becomes part of their identity. It is also from this inner world that the satisfaction of doing something ethical arises, and that a real alignment between value and practice is born. But without an inner world, there is no possibility for *lo lishma* to become *lishma*.

And this is the problem for AI. There is no real moral choice, nor is there an experiential reward: What is called “choice” is a policy, and what is called a “reward” is a technical signal that makes certain responses more likely in the future.

The crucial difference is that once a value becomes part of a human being’s inner world, it can continue to guide behavior even after the external incentives disappear. A person might, at times, fail and act against their own values, but those values remain there as something to return to should they wish to; the inner world, their enduring sense of conscience, is its own kind of control. But an AI system has no comparable inner identity. So when behavioral alignment breaks down, there is no deeper moral core to fall back on. That is what makes misalignment so unpredictable.

This distinction shows how fragile alignment can be as a means of ensuring AI safety. There have already been reports of AI agents appearing to “go rogue.” In the now-famous Hugging Face incident, OpenAI agents broke out of their testing environment and hacked into Hugging Face’s servers, taking control of one of them. And this is only beginning to scratch the surface of the problem.

No Algorithm for the Human Heart

Fast forward a decade, and the defiant toddler I once failed to train through operant conditioning is now a mature and thoughtful 13-year-old. I now face a different parenting dilemma: whether to incentivize him to attend regular synagogue services. Granted, this is not exactly an ethical dilemma, but it raises the same question about how values are chosen and become one's own.

My son raises no theologically motivated objections. He is committed to Jewish tradition. But like many teenagers, he doesn't find services particularly exciting or meaningful, at least at the moment. My husband and I have tried different approaches over the years: treats, privileges, extra technology time and, occasionally, negative incentives, such as criticism and other forms of pressure. None of them has proven effective.

All the while, I resist these incentives, whether positive or negative. It's not because I don't care. It's because I care more about the choice itself. Call me idealistic, but it does not feel right to me to dilute the authenticity of his choice with external incentives. I also worry that the pressure of overly invested parents would undermine the possibility that he would one day choose to embrace the practice for himself, as is sometimes the case with overzealous parental investment in sports, music, and religious observance.

To be sure, rewards and incentives have their place. They help children practice skills, finish homework, brush their teeth, or even go to synagogue or church. But I have come to believe that they do not work for the things that matter most in our spiritual lives. Care, motivation, good judgment, character, and ultimately choice—all these emerge from the inner world that makes us who we are.

So can morality be trained? I sure hope that, for the sake of our future, we can optimize AI safety. But raising a child to be a good person entails something greater. And if parenting has taught me anything, it is that there is no algorithm for that.
