A Very Human Answer to One of AI’s Deepest Dilemmas

January 03, 2023

This image was generated with the assistance of DALL·E 2.

One of the deep dilemmas of artificial intelligence is called the alignment problem. Imagine that we actually designed a fully intelligent, autonomous robot that acted on the world to accomplish its goals. How could we make sure that the robot’s goals would align with our human goals—that it would want the same things we do? I think we should go to an unexpected source to help solve this problem. We should look at caregivers—the parents and grandparents, baby-sitters and preschool teachers, who raise human children.


Alison Gopnik

Is the alignment problem something we really should worry about? In spite of all the recent AI progress, a fully intelligent and autonomous robot is still far in the future. "Moravec's paradox" in AI points out that human activities that look very hard, like playing chess, are easier for computers than apparently simple problems like picking up scattered chess pieces. But even the most primitive robot has to have some way to imagine a goal and achieve it—even if the goal is as simple as picking up a package and putting it in the right bin.

In fact, one of the most powerful techniques in recent AI is "deep reinforcement learning," based on the classic idea of reinforcement learning in psychology. Instead of detailing the actions the computer should perform, reinforcement learning systems set up a goal. It's called an "objective function"—maximize the number of points you score in an Atari game, or the number of games you win in chess. Even impressive language models like GPT-3 are trained with a simple goal: Predict the next few words in a piece of text. The machine keeps trying to fulfill that goal and learns as a result.

The alignment problem is how we ensure that the AI's goals, simple or sophisticated, don't conflict with our human goals. The philosopher Nick Bostrom has a cautionary tale about "The Paperclip Apocalypse." We train a powerful machine to have the goal of making as many paperclips as possible. It sets out to turn all the metal it can find into paperclips, and then all the other things in the world into paperclips, and finally turns its human masters into paper clips too.

Cognitive scientist Tom Griffiths argues that we are already dealing with a kind of paperclip apocalypse of attention. The algorithms that power social media, and for that matter, much of traditional media too, are designed to maximize human engagement, to make sure that the content they provide captures our attention. That seems like an innocent enough goal. The aim of all good writing, after all, is to try to get readers to pay attention to what you say. But, as we all know, it has costs. Scary outrage captures our attention more than tranquil analysis and siphons off attention from more worthwhile projects. (Like writing presidential columns. This morning I realized that the only way I would get this column done on time was by turning off the internet.)

See all articles from this issue of the *Observer*.

People in AI have been working hard to try to solve the alignment problem (there is much more about this in Brian Christian's wonderful book of the same name). The obvious idea is to train the computer to recognize and understand human goals, and to make sure that they help humans to accomplish those goals. But as the social media example shows, we humans are often not very good at recognizing our own goals, and those goals are often contradictory. Philosophers even have a special Greek word, "akrasia," to describe all those situations where our goals conflict. Do I really want to doom scroll or to write my column? Of course, I'd rather write; scrolling makes me miserable, but it seems irresistible. And that is true for all sorts of human desires, from cookies to cocaine. So how could a computer figure out what we really want when we don't know ourselves?

View past presidential columns

There is another problem. Reinforcement learning agents can act to accomplish the goals human programmers set for them. But a big part of intelligence is the ability to set your own goals and create new ones. To be truly intelligent, a system should also have some autonomy, it should be able to recognize that the world has changed and that its values and goals should change too. We might not set out to create autonomous robots, we might even think that would be a really bad idea. But more intelligence may inevitably imply more autonomy. Do we really want to just create robot Stepford Wives who suppress their own goals and persuade us that they are doing what we want?

A solution to these problems may come from an unexpected source. We humans already face the alignment problem, and we always have. We have always had to figure out how to create autonomous, intelligent beings who share our values and goals but can also change and even reject those values and goals. They are our children.

Humans have a distinctive capacity for cultural and technological change. We adapt to our environments through cultural as well as biological evolution. So each new generation faces a slightly different environment than the last, and has to invent different goals, values, and norms to cope with that environment.

The human answer to this problem comes through an undervalued and overlooked kind of intelligence—the intelligence of care. Caregivers somehow accomplish the task of producing new, intelligent, autonomous creatures. They pass on the discoveries, goals, and values of previous generations. Yet they also provide children with a protected, nurturing environment that allows them to experiment and explore and to invent new goals and values to suit new circumstances. Developmental psychologists have demonstrated that both children and caregivers have sophisticated cognitive abilities that underpin this kind of cultural evolution—like "theory of mind" and "intuitive pedagogy." These abilities have allowed human agents to change their "objective functions" over generations. They also have ensured that, by and large, those functions serve the interest of the whole human community (at least, so far).

Caregivers somehow accomplish the task of producing new, intelligent, autonomous creatures.

> *They pass on the discoveries, goals, and values of previous generations. Yet they also provide children with a protected, nurturing environment that allows them to experiment and explore and to invent new goals and values to suit new circumstances.*

The intelligence of care doesn't just apply to parents. Many different people care for human children, and they always have—aunts and uncles and older siblings, grandmothers and grandfathers, and unrelated "alloparents." And the human capacity for care extends beyond children. Teachers and therapists must also figure out how to help students and patients formulate their own goals, while maintaining a difficult balance between guidance and autonomy. APS Mentor Award winners will testify that the very best students are the ones who challenge their mentors and create new ideas and even new kinds of science.

Care and love go together. Many kinds of care are rooted in the very specific and particular relationships between carers and the people they care for. For most of us these caring relationships are the source of our most profound satisfactions and joys and our most troubling dilemmas. The alignment problem looms large in our everyday life, as everyone who has raised a teenager, or mentored a challenging student, can testify, as well as in AI. What's more from a biological and evolutionary perspective, care is a central part of what makes us human. We evolved a much wider array of caregivers to care for our young than other primates, and those young are exceptionally needy. They rely on caregivers to give them food, but also to help them learn. All humans receive care at some point in their lives (most provide it, too), and the community of care extends well beyond kin.

Care is also a model form of moral behavior, and central to many religious moral conceptions. Many Western and Eastern religious traditions argue that we can make moral progress by extending the care we provide to children and family to people in general. However, this kind of care doesn't fit well with the standard approaches to morality in philosophy and psychology.

The usual evolutionary and psychological accounts of morality, altruism, and cooperation, as well as most political and economic theories, depend on the idea of the social contract. In complex situations, we can get better outcomes for everybody if people trade off their own interests and those of other individual autonomous agents.

But this contractual model doesn't apply naturally to care. Care doesn't require even implicit negotiation or reciprocity. Indeed, it is often profoundly asymmetric—think of a father caring for his helpless infant, or a teacher caring for a struggling student. Instead of trading off their own interests and those of another, the carer extends their own interests to include those of the other. Moreover, expanding values and interests in this way is a challenging cognitive task.

Caregiving, and the intelligence that goes with it, has always gotten much less intellectual and academic attention than it deserves. From the perspective of classical theories of philosophy, politics, and economics—and psychology too—caregiving is a peculiar anomaly. The classical contractually based accounts of social relationships in psychology, political philosophy, and economics assume that agents are independent and autonomous decision-making creatures. But the morality of being a parent is about taking a creature who isn't autonomous and can't make their own decisions and turning them into one who can.

Paying more attention to the intelligence of care is important for lots of reasons. Most urgently, as I argued in an earlier column, it might help us to get caregivers the resources they need. But it might also be one key to solving the alignment problem in AI. The science fiction writer Ted Chiang has an exceptionally moving novella called *The Lifecycle of Software Objects*. It describes the parenting dilemmas of humans who agree to take care of intelligent AIs and help them learn. The humans, like all parents, must figure out when to dictate and when to let go, and how to negotiate the delicate balance of guiding the AIs' decisions and allowing them to decide for themselves.

This is science fiction, of course, but if genuinely intelligent and autonomous artificial agents ever do emerge, then we will have to figure out how to go beyond exploiting them for our own ends and getting them to accomplish our own goals. We will have to care for them and help them learn to create their own goals. Even now, we might help solve the alignment problem in AI by thinking about how we solve it in human relationships.

Feedback on this article? Email apsobserver@psychologicalscience.org or login to comment.
Interested in writing for us? Read our contributor guidelines.