Computational Frameworks for Human Care

Brian Christian

Some of the earliest science-fiction literature to imagine humans' long-term relationship with machines portrayed technology as a kind of caregiver for humans. The retrofuturist vision of machine care is poised to become reality, as the world of the 2020s is experiencing both incredible advances in AI technology as well as demographic changes that, together, make such systems seem at once possible and necessary. Tracing the key themes from our literary and cultural imagination and framing them against the technical progress in the field of AI alignment reveals insights and lessons for us as we approach the prospect of bringing certain forms of computational care to life. In so doing, they provide not only practical guidance but also give us an opportunity to sharpen our intuitions about the nature of human care itself.

In 1909, E. M. Forster published his uncannily prescient story "The Machine Stops," portraying a future in which humans live solitary lives in small apartments, interacting with one another holographically, all of their needs for food and sleep provided by the "Machine."[1]

The Machine has an operating manual, called the book of the Machine. "If she was hot or cold or dyspeptic or at a loss for a word," Forster explains of the main character, Vashti, "she went to the book, and it told her which button to press." Sometimes even this minimal effort is not required. At one point during a video call with her son, Vashti says she is feeling unwell: "Immediately an enormous apparatus fell on to her out of the ceiling, a thermometer was automatically laid upon her heart. She lay powerless. Cool pads soothed her forehead.... Vashti drank the medicine that the doctor projected into her mouth, and the machinery retired into the ceiling."

Not only does the Machine embody a certain form of caregiving – admittedly a sometimes overzealous one – but Forster is explicit that it has supplanted human forms of care. "'Parents, duties of,' said the book of the Machine, 'cease at the moment of birth.'"

It is striking how central the theme of mechanical caregiving is to the fictive imagination of humans' technological future. Fast-forwarding to the mid-twentieth

https://doi.org/10.1162/daed_a_02131

century – the Golden Age of Science Fiction that accompanied, rather than anticipated, the birth of the computer – we see a similar vision. In Ray Bradbury's classic 1950 story "The Veldt," we meet the Hadley family, who live in a "Happylife Home . . . which clothed and fed and rocked them to sleep and played and sang and was good to them."[2] (Something in the breathless polysyndeton suggests a kind of inexhaustibility that is, in fact, exhausting.) We see how the house itself has taken over the managing of all domestic activities; we see the mother, Lydia, watching "the stove busy humming to itself, making supper for four."[3]

Indeed, the house has supplanted the parents' care, not only for their children, but also for each other: "The house is wife and mother now, and nursemaid."[4] The gendered language here is striking; it feminizes the Home while reflecting now-dated 1950s norms, and in doing so suggests something very particular about the house. It hasn't supplanted the need for an economically productive breadwinner (on the contrary, Bradbury emphasizes its monetary cost), but it *has* replaced human labor in its myriad other, non-GDP-denominated senses: foremost, care.

By the late 1960s, the world had entered the era of manned spaceflight, and the poet Richard Brautigan, as a writer-in-residence at Caltech, wrote memorably of "a cybernetic ecology / where we are free of our labors / [. . .] / and all watched over / by machines of loving grace."[5] Part of the poem's lasting appeal is its inscrutable tone: Is it earnest? Ironic? Sinister? It resonates all the more for this ambiguity, which speaks to our own ambivalence about what the long-term future holds, and even what it should.

But this vision of AI as the ultimate caregiver for humankind is – critically – hardly exclusive to the arts. Rather, the idea of machine caregiving has been a guiding light for the field of artificial intelligence itself. Caregiving is often cited as part of the teleology of AI: one of the canonical use cases of intelligent machines, one of the primary things that intelligent machines are *for*. As the field has progressed, particularly with the resurgence of artificial neural networks since 2012 and the rise of large language models (LLMs) since 2019, a version of this goal is starting to come within reach. Meanwhile, demographic changes are leading to a critical shortfall in caregivers, a shortfall that political leaders view as impossible for human labor to fill. Supply and demand, in other words, are poised to meet.[6]

An authoritative accounting of this progression is beyond the scope of this essay, but a somewhat arbitrary sample is sufficient to indicate its size, shape, and velocity: "The biggest argument for robot caregivers," argued gerontologist and author Louise Aronson in a *New York Times* op-ed in 2014, "is that we need them. We do not have anywhere near enough human caregivers for the growing number of older Americans."[7] In 2015, a Canadian long-term care facility began a pilot program with a robot that could play bingo with residents.[8] By 2017, 59 percent of Americans viewed the idea of robot caregivers as "realistic."[9] In 2018, the government of Japan was predicting a shortfall of 370,000 caregivers by the year 2025 and had al-

ready spent over $300 million in research funding toward care robots.[10] In early 2022, a series of nursing homes in Minnesota, beginning with the Estates of Roseville, began introducing care robots for patients with Alzheimer's and dementia.[11] Later that year, the U.S. National Institutes of Health awarded a $2.8 million grant to researchers at the University of New Hampshire to research care robots.[12] In 2023, *The New York Times* reported on a pilot program for deploying robots in both care facilities and individual homes in "Italy's most innovative region for elder care."[13]

Meanwhile, as the frontier of AI capabilities has been dramatically advancing over the past ten to fifteen years, we are seeing the AI-research community engage more and more explicitly with normative questions of ethics, safety, and control: in particular, how to impart human values into AI systems in the kind of numerical form that a machine-learning system can understand and pursue. This question has come to be known as "the alignment problem," and its corresponding subfield of AI research has come to be known as "AI alignment."[14]

To what extent does the conceptual structure used by the alignment research community resemble a notion of care that might be recognizable to another field?

Though the subject of caregiving seems at a glance to be quite disparate from the field of artificial intelligence, the two were bound together from before the birth of the computer and are in an ever-closer relationship now. So let me begin by unpacking the actual computational mechanisms of the systems being built and deployed today. What we will find is that as the computational techniques for designing and training AI systems change, so does the implied *relationship* between the system and its designer or user. Furthermore, not only is this relationship evolving, but it is moving toward a computational articulation of a *caregiving* relationship. Progress in that direction is considerable – but the differences and gaps are just as striking.

Contemporary AI systems are rooted in a branch of computer science known as reinforcement learning (RL), which began in the late 1970s and early 1980s. RL itself draws heavily from the behaviorist tradition in psychology and is concerned with the maximization of numerical "rewards."[15] In the RL conception, an "agent" exists in an "environment" and learns a "policy" for taking actions that transition it within "states" of that environment. Typically, the agent begins with a randomized policy and learns through some form of trial and error to take actions that maximize the expected value (or discounted expected value) of future rewards. Fundamental to this conception is the "reward function," which is a mapping of combinations of states and actions to scalar-valued rewards. In effect, the system treats its environment as a kind of video game in which it is trying to score points.

The RL framework has been responsible for a number of signature successes in the field of AI, perhaps most notably, and fittingly, in game-playing itself: RL sys-

tems came to master the game of backgammon in the 1990s and the game of Go in the 2010s, culminating in the defeat by DeepMind's AlphaGo system of legendary Go champion Lee Sedol in 2016, followed by the number-one-rated Go player at the time, Ke Jie, in 2017.[16]

RL takes the problem of directly writing *code* to pursue a person's objectives with the problem of writing a *reward function* that will incentivize an agent to do what they want. It thus frames the human as a "reward designer," and the relationship it imagines, and engenders, is of a superior creating incentives and a subordinate following them. It is up to the human to express their desires or needs as a mathematical object – the reward function – and to design reward functions that fully capture those desires.

In practice, reward design is very challenging, and RL researchers are accustomed to discovering, time and again, that their system is exploiting some loophole in their specified reward function: namely, doing what they asked for, but not what they meant. For instance, in their graduate student days in the 1990s, Google's David Andre and Astro Teller built an RL system to play soccer, and in order to incentivize the system to learn how to score goals, they gave it an incentive – worth a fraction of a goal – for taking possession of the ball. The system learned to approach the ball, vibrate its paddle, and "take possession" of the ball many times per second.[17] In 2016, OpenAI researcher Dario Amodei was training an RL agent to play a boat racing game called Coast Runners; as it would be too complex to directly encode a notion of track position, laps, and placement relative to the other boats, he gave it the more straightforward objective of maximizing in-game points. Amodei believed this would correspond to good racing, but the system learned to quickly veer off of the track into a harbor that contained a replenishing supply of power-up items, where it would drive in haphazard circles, forever.[18] These examples are par for the course in RL and are a significant part of why the AI-safety community has come to view RL as dangerous.

It's also worth reflecting on the role of incentives in a care relationship. Parenting, for instance, does involve a lot of incentive design, both consciously and unconsciously – and it even has some of the same failure modes. Economist Joshua Gans decided to give his daughter a dollar every time she helped her younger brother use the toilet; he soon discovered she was force-feeding her sibling water in order to make as much money as possible.[19] Cognitive scientist Tom Griffiths praised his daughter for sweeping up crumbs with a brush and dustpan; she then dumped the dustpan out onto the floor, in order to sweep them up again and get a second helping of praise.[20] These are, at the mathematical level, essentially identical to the "reward cycles" of the soccer robot and the boat above. Notice, however, that in parenting, it is the care*giver* who designs the reward structures, and when they do, it is generally a short-term choice that serves their broader aims. It is a tactic, in other words, not a strategy – and certainly not the foundation of care.

We have seen how manually designing a reward function can lead to loopholes and unexpected behavior, but this is not the only drawback of reinforcement learning. What *is* the reward function that best corresponds to winning a boat race? Or keeping a home tidy? For all its successes, RL does not offer us a way forward in the cases in which we cannot easily formulate an explicit mathematical representation of the behavior we want.

The next major step forward for the field of AI, around the turn of the millennium, was to turn the discovery of a proper reward function for a task into – itself – a problem for the machine to solve. From this insight came the technique known as *inverse* reinforcement learning (IRL). If a reinforcement-learning agent is tasked with finding the right set of behaviors (the right "policy") to maximize a given reward function, then inverse reinforcement learning goes the other direction: given a policy – in this case, observations of human behavior – can the agent infer the reward function the human appears to be pursuing?[21] If so, then the agent can adopt that reward function as its own (and use standard RL to translate that reward incentive into actions of its own).

By the late 2000s, this idea had proven itself in a series of successes, perhaps most dramatically the "Stanford helicopter" work by a team of computer scientists including then doctoral student Pieter Abbeel and his advisor Andrew Ng to design a radio-controlled (RC) helicopter that would train itself to perform complex autonomous stunts. It would do this by observing demonstrations provided by human RC pilots and then inferring numerical "reward models" to capture what those pilots were attempting to do. In other words, it would turn stunts into math. Having a numerical description of a given stunt, it could then use standard RL techniques to learn the set of actual behaviors – the torques and accelerations and corrections – that would enable it to perform that stunt.[22] Subsequent work has shown AI systems capable of inferring numerical reward functions to describe everything from taxi-driving to the act of putting dishes in a dish rack.[23] Again, this makes for a significant extension of AI – into domains where we can directly show what behavior we want without needing to specify it in numerical terms.

There are fundamental connections between the computer-science literature on IRL and various concepts in developmental cognition that suggest that we humans have some kind of innate "IRL" capacity and drive. For instance, psychologist Felix Warneken studies the developmental roots of altruism, cooperation, and helping behavior in humans, and has demonstrated quite strikingly that humans possess an intrinsic ability to infer the goals of others and desire to help others achieve those goals. Children as young as eighteen months of age can, for example, observe an adult trying to pick something up out of reach or open a door with their hands full, and the children will spontaneously help.[24] (Notably, this is multiple years before they are able to pass the "Sally–Anne test," which suggests that inferring others' *goals* happens significantly earlier in development than in-

ferring their *beliefs*.)[25] The instinctive human helping behavior also, even by eighteen months, appears to exceed the capacities of our nearest primate kin.[26] So it would seem that this impressive ability to infer others' goals, and the corresponding drive to pitch in, is more or less hardwired and nearly unique among the animal kingdom.

IRL transforms the implicit relationship of human and machine once again: from manager and subordinate to something more like teacher and apprentice.[27] Here we begin to see certain aspects of caregiving beginning to formally enter into the technical AI research literature. An IRL system, as we've seen, rather than being handed an explicit objective, begins by observing its human user, then infers the objective the human appears to be pursuing, and finally adopts that inferred objective as its own. There are echoes here of several of the other essays in this volume. For instance, Eric Schwitzgebel discusses the idea of our concern for others as being rooted in an empathic reaction of coming "to want or loathe what they want or loathe."[28] Ashley J. Thomas and colleagues describe how aspects of the relationship between a carer's goals and the cared-for person's goals can represent a reversal of a normal power relationship: namely, instead of the more powerful and capable individual subordinating the other to carry out their own goals, they do the opposite and assign the goals of the less capable, cared-for person to themselves.[29]

This is indeed an important dimension of care, and IRL does seem to capture aspects of a caregiving relationship. There are nontrivial technical challenges, to be sure: for instance, "indexical" issues where we must take care that the AI system has the correct frame of reference when inferring the reward function to pursue. If it sees me reaching for a coffee cup, it should pursue the reward function for getting *me* the coffee, not obtaining the coffee for *itself*.[30] Handled correctly, the IRL framework allows us to imagine a domestic robot, for example, that can do approximately what an eighteen-month-old can do: see us reaching for an object beyond our grasp and hand it to us, or see us approaching a door with our hands full and open the door. In elder care, we could imagine such a system helping a human to stand up, to traverse to the bathroom, and so forth.

With this said, IRL – and more broadly the adoption of a cared-for person's goals as the caregiver's own – does not by itself fully constitute what we would want or expect from a caregiving relationship. To start, IRL is by default limited to the things that we ourselves can demonstrate, even if imperfectly. For a child with limited motor skills, or an elder with limited mobility, this might pose a challenge. Second, there are many aspects of care that involve providing help in ways not explicitly asked for or modeled. Finally, caregiving – especially for children – often requires *denying* explicit requests and physically intervening *against* behavior, rather than facilitating it. What reward function does this sort of caregiving behavior pursue?

The current state of play in AI alignment research can be said to have begun with a 2017 collaboration between DeepMind and OpenAI, then the world's two premier AI research labs, centered on the question of how to get a simulated bipedal robot to perform a backflip.[31] At first glance, this was simply the helicopter project with a different form factor. But there was a subtle, and crucial, difference: while the Stanford helicopter project used expert RC pilots to supply the demonstration data, it's nearly impossible to get a bipedal robot to perform a backflip using buttons and joysticks. (And most people can't do a backflip, so that form of demonstration was also out.) Despite the fact that people cannot specify a backflip directly in numerical terms, nor can they demonstrate one, they can nevertheless immediately recognize a backflip when they see it.

Might that be enough?

The system would begin by wriggling around at random, and then present the user with two video clips and ask them which they preferred: which was infinitesimally closer to what the user had in mind? The user would select one of the two clips, and the process would repeat. After just a few hundred of these comparisons, over the span of about an hour, the robot would be doing beautiful, picture-perfect backflips and sticking the landing.

This procedure has come to be known as "reinforcement learning from human feedback," or RLHF. OpenAI wasted little time in transferring this methodology from a kinesthetic domain to a linguistic one. Soon they were asking crowdworkers which of two passages was a better summary of a document, or which of two answers to a question they preferred.[32] Crowd-worker preferences are used to build a "reward model" that assigns numerical rewards to language outputs, and that reward model is used in turn to create a text-dialog system that learns to generate responses consistently rated highly by the reward model. This is the process behind the breakthrough success of ChatGPT and the many LLMs that have followed in its wake.[33]

RLHF once again shifts the relationship between human and machine. Compared to IRL, which takes a roughly "second-person perspective," adopting the user's goals directly, RLHF can be thought of as taking a roughly "third-person perspective": it presents the response that would be maximally approved of by a focus group. This is a sort of *democratic* notion of care, for better or worse.[34] Indeed the question of whose values – whose reward function – these systems embody has become central to this technology and is likely to remain so. The exact degree of input that is appropriate from states, the companies that build these systems, the third-party raters, and the individual users themselves is not obvious.

Having established, broadly, the mathematical foundations of human-AI interaction as they stand in the mid-2020s, at the precipice of a broad deployment of caregiving technologies – large and small, physical and virtual – we are now in

a position to consider the conceptual issues that will shape how these systems behave – and how they ought to.

I opened this discussion with points of reference in the science-fiction canon, and it is worth revisiting those texts with the framework of AI alignment now more firmly in mind. Remembering that both Forster and Bradbury present us with what are essentially *cautionary tales*, we can use the Machine and the Happylife Home as foils, and, in the context of real-world AI alignment, see what they reveal to us about a normative account of machine care. The first of these themes is the combination of acceptance and empowerment.

In "The Machine Stops," Forster's humans come to think of the Machine in terms that range from the parental to the divine: it "feeds us and clothes us and houses us; through it we speak to one another, through it we see one another, in it we have our being." Yet all is not well – not at all – with the kind of care that the Machine provides. For one thing, we come to learn that "Each infant was examined at birth, and all who promised undue strength were destroyed." Forster's narrator describes this as if it were regrettable but necessary: "Humanitarians may protest, but it would have been no true kindness to let an athlete live; he would never have been happy in that state of life to which the Machine had called him; he would have yearned for trees to climb, rivers to bathe in, meadows and hills against which he might measure his body." Here is perhaps Forster's first lesson for us about the nature of care. Care requires the caregiver to accept the cared-for *as they are*. There is very clearly something wrong with a caregiver *killing* someone for whom their style of caregiving would not be helpful.

Tragically, the human relationships in "The Machine Stops" suffer from precisely this same fault. That Vashti's son Kuno needs her to visit him *in person* is our first clue. He wishes to be seen (both literally and figuratively): to be understood, accepted, recognized, not judged. He understands that he cannot get this from the Machine, nor from his holographic interactions, including with Vashti herself. Unfortunately, we learn that he cannot get it from Vashti either. Her empathy is limited, and her visit brief. By existing in a world in which at least a certain category of human needs is so routinely and automatically met, they have lost a core part of their humanity: the ability to support one another.

Over the course of the story, Kuno radicalizes. "Cannot you see," he says, "that down here the only thing that really lives is the Machine?" It caters to human desires in an immediate sense, but the shape of their lives, the nature of their relationships to each other, their sense of imagination and of what is possible, are all confined within the terms the Machine sets. That is not care.

True care must include, crucially, empowering people to care for each other, and also to no longer *need* care. In this, Forster's Machine is a failure. It caters to a subset of needs while fundamentally disempowering people: from caring for each

other, from caring for themselves, and most of all from a life independent of the Machine itself. "We created the Machine, to do our will, but we cannot make it do our will now," Kuno says. "The Machine develops – but not on our lines. The Machine proceeds – but not to our goal."

Likewise in Bradbury's vision, the family's relationship with the Happylife Home oversteps the mark when it disempowers them as caregivers for one another. George and Lydia hire a psychologist to assess them, and he admonishes them: "You've let this room and this house replace you and your wife in your children's affections. This room is their mother and father, far more important in their lives than their real parents."[35] We have seen how, in children, the impulse to help others is a deeply rooted one, present almost from birth. This, it would seem, is one of the few needs that neither the Machine nor the Happylife Home can provide.

The Home is described in terms that are sometimes inspiring: "the nursery caught the telepathic emanations of the children's minds and created life to fill their every desire."[36] But more often, we see its effect on the adults and children alike as enfeebling. When the father, George, announces that he plans to turn the machinery off for a period, the children rebel: "That sounds dreadful! Would I have to tie my own shoes instead of letting the shoe tier do it? And brush my own teeth and comb my hair and give myself a bath?"[37]

In the real world, we often hear techno-optimists arguing that humans can use the time and energy that future AI systems will free up from errands such as these to pursue intrinsically meaningful activities like the arts. In Bradbury's conception, however, the Home's enfeeblement comes equally to the arts. "I didn't like it when you took out the picture painter last month," says the son, Peter. "That's because I wanted you to learn to paint all by yourself, son," George replies. "I don't want to do anything but look and listen and smell; what else *is* there to do?"[38] The machine has reduced participation to passive consumption.

The celebrated Lebanese-American poet Kahlil Gibran, in his book *The Prophet*, uses the metaphor of an archer to describe parenthood: "You are the bows from which your children as living arrows are sent forth."[39] It is a process of preparing the child to be free and self-sufficient. Indeed, parenthood involves, as psychologist Nim Tottenham puts it, a "seeming paradox: initial dependence gives rise to independence."[40] Of course, later-in-life care, and in particular hospice care, cannot have this exact character. But it still retains something of its spirit: to the extent possible, the caregiver prepares the cared-for person for an experience of their own, whether that experience is early adulthood or death. The caregiving relationship may be good in itself, but it is not an end in itself.

Ironically, the second theme of caregiving exists in a slightly paradoxical tension with the first. Despite the fact that caregiving requires us to see and accept the cared-for person on their own terms, and to empower them to pursue

their goals, including the goal of no longer needing our care, caregiving, especially of children, is *not* carte blanche. Indeed, sometimes the most caring thing a parent can do for their child is to physically intervene between the child and the object of their desire – or simply to say *no*.

On what basis can this be justified?

Perhaps nothing in science fiction more memorably embodies the horror of a machine-human relationship breaking down than the moment in *2001: A Space Odyssey* when HAL denies Dave's request to open the doors to let him back on the ship: "I'm sorry, Dave, I'm afraid I can't do that."[41] And yet Bradbury points out that there are horrors of acquiescence, too. The children in "The Veldt" become increasingly disturbed and moody, and the virtual world their nursery creates for them amplifies, rather than mollifies, this darkness. Their play space becomes something violent and eerie: namely, the titular veldt, in which lions and vultures feed on flesh while screams echo in the distance. The psychologist that George and Lydia hire is instantly concerned, concluding that "the room has become a channel toward – destructive thoughts, instead of a release away from them."[42] It is a line that feels discomfitingly allegorical for any twenty-first-century users of recommender systems and social media.

By default, the "revealed preferences" or "reward function" of a compulsive gambler, say, or a compulsive shopper, put an IRL system in the position of an *enabler*. On what basis, then, do we – and might a machine – *deny* the apparent wants of a human user?

There are a number of theories that are perhaps best explored in the philosophy literature, but we can easily enumerate several candidates.

Perhaps we, as parents, simply have our *own* reward function, which can conflict with the assumed goals of the child's. In other words, our desire to have our child not be electrocuted overrides our child's desire to put a metal object into an electrical socket. RLHF embodies a certain communal (if hegemonic) form of this: today's language models will typically decline to assist a user who wants to build a bomb, extort a coworker, commit fraud, or anything else that violates the preferences of certain others (be it the state, the company, and/or the focus group who provided preference data).

Perhaps we assume not only the child's *present* goals but some notion of the goals of their *future self*; surely the adult our child will become will be grateful we didn't let them electrocute themselves as a child, and they might even be grateful that we limited their candy intake and screen time. We have seen stirrings of this sort of movement in critical perspectives on technology: for instance, the Time Well Spent movement of 2016, which encouraged social-media companies to optimize for the *retrospective* preferences rather than in-the-moment impulses of their users.[43]

Perhaps we have some way of understanding that human goals are sometimes in conflict with one another, and as carers, we aspire to serve the "higher" purposes.

AI researchers are beginning to imagine ways of approaching these ideas within the context of reward modeling and alignment.[44]

And perhaps we have some more objective notion of well-being – we care not only about what you *want* but what's *good for you*. Neuroscientist Kent Berridge, for instance, has shown that "wanting" and "liking" comprise two distinct reward systems in the brain.[45] It's not clear which form of rewards social-media companies, for instance, are even *trying* to cater to. There is clearly an information asymmetry to be overcome here. From the perspective of, say, developers at Netflix, they have a wealth of data about what people will click and how long they will watch it. It's much less clear whether a late-night TV binge was *good* or *bad* for them – either in their own retrospective opinion or according to some more objective metrics. But true care aspires to go beyond the cared-for's needs in the moment, and so should our machine helpers, even at their most quotidian.

We have seen, in sum, how deeply ingrained the notion of machine care is, not only in the science-fiction imagination – where it ranges from the utopian to the horrific – but also in the aspirations of the field of artificial intelligence, and in the minds of policymakers looking to artificial intelligence for solutions to a future crisis-level shortfall in care workers.[46] As AI alignment research has progressed by stages – from code to rewards to demonstrations to preferences – so has the relationship that increasingly pervasive AI systems have with their human designers and human users. This progression has come to resemble, at least in certain dimensions, a relationship of care – but there is much to be desired, many open problems to be addressed, and many normative questions to be considered.

It is often said that we don't fully understand something until we've taught it to someone else; indeed, the very act of teaching something is often an important last step in distilling or sharpening our own inchoate knowledge. The prospect of machine care is just such an opportunity. As is so often the case, the process of trying to formalize core aspects of the human experience is revealing to us what care really *is* – and perhaps even how much we have yet to understand about it. Let us take this moment, then, as an opportunity – if a somewhat urgent one – to confront and explore just what it means to care and be cared for, including by one another.

ABOUT THE AUTHOR

Brian Christian is an author and researcher whose work explores the human implications of computer science. He is affiliated with the Human Information Processing Lab at the University of Oxford, the AI Policy and Governance Working Group at the Institute for Advanced Study in Princeton, N.J., and the Center for Information Technology Research in the Interest of Society and the Center for Human-Compatible AI at the University of California, Berkeley. He is the author of *The Most Human Human* (2011), *Algorithms to Live By* (with Tom Griffiths, 2016), and *The Alignment Problem* (2020).

ENDNOTES

[1] E. M. Forster, "The Machine Stops," in *The Eternal Moment and Other Stories* (Harcourt, Brace and Company, Inc., 1928), Project Gutenberg, "The Eternal Moment and Other Stories," February 7, 2024, https://www.gutenberg.org/cache/epub/72890/pg72890-images.html#the_machine_stops.

[2] Ray Bradbury, "The Veldt," first published as "The World the Children Made," *The Saturday Evening Post*, September 23, 1950, and later in Ray Bradbury, *The Illustrated Man* (Bantam Books, 1967), 7.

[3] Ibid.

[4] Ibid., 10.

[5] Richard Brautigan, "All Watched Over by Machines of Loving Grace," in *All Watched Over by Machines of Loving Grace* (The Communication Company, 1967).

[6] For more on the growing demand for technological solutions to care, and reasons for caution, see Elizabeth Fetterolf, Andrew Elder, Margaret Levi, and Ranak B. Trivedi, "Technology and the Dynamics of Care for Older People," *Dædalus* 154 (1) (Winter 2025): 117–133, https://www.amacad.org/daedalus/technology-dynamics-care-older-people.

[7] Louise Aronson, "The Future of Robot Caregivers," *The New York Times*, July 19, 2014.

[8] Mercy Cuttler, "Robot Caregivers Aim to Improve Seniors' Quality of Life," CBC News, January 21, 2015.

[9] Aaron Smith and Monica Anderson, "Americans' Attitude toward Robot Caregivers," in *Automation in Everyday Life* (Pew Research Center, 2017), 39–48.

[10] Daniel Hurst, "Japan Lays Groundwork for Boom in Robot Carers," *The Guardian*, February 6, 2018; and James Wright, "Inside Japan's Long Experiment in Automating Elder Care," *MIT Technology Review*, January 9, 2023. See also James Wright, *Robots Won't Save Japan: An Ethnography of Eldercare Automation* (ILR Press, 2023).

[11] Eric Hal Schwartz, "Minnesota Nursing Homes Introduces [*sic*] Robot Caregivers," *Voicebot.ai*, January 7, 2022.

[12] Simon Spichak, "Dementia Companion: Robo Caregivers Are Coming to Help," *Being Patient*, September 13, 2022.

[13] Jason Horowitz, "Who Will Take Care of Italy's Older People? Robots, Maybe," *The New York Times*, March 25, 2023.

[14] For a history and overview, see Brian Christian, *The Alignment Problem* (W. W. Norton & Company, 2020).

[15] Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction* (The MIT Press, 2018).

[16] For backgammon, see Gerald Tesauro, "TD-Gammon, A Self-Teaching Backgammon Program, Achieves Master-Level Play," AAAI Technical Report FS-93-02 (Association for the Advancement of Artificial Intelligence, 1993). For AlphaGo, see David Silver, Aja Huang, Chris J. Maddison, et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature* 529 (7587) (2016): 484–489.

[17] This anecdote is related in Andrew Y. Ng, Daishi Harada, and Stuart Russell, "Policy Invariance under Reward Transformations: Theory and Application to Reward Shaping," *ICML* 99 (1999): 278–287. The soccer project itself is described in David Andre and Astro Teller, "Evolving Team Darwin United," *RoboCup-98* (Springer, 1999).

[18] Jack Clark and Dario Amodei, "Faulty Reward Functions in the Wild," OpenAI Blog December 21, 2016, https://openai.com/index/faulty-reward-functions.

[19] See Chana Joffe-Walt, "Allowance Economics: Candy, Taxes and Potty Training," NPR, September 3, 2010; and Joshua Gans, *Parentonomics: An Economist Dad Looks at Parenting* (The MIT Press, 2009).

[20] Tom Griffiths, personal interview, June 13, 2018.

[21] Stuart Russell, "Learning Agents for Uncertain Environments (extended abstract)," *Proceedings of the Eleventh Annual Conference on Computational Learning Theory* (ACM Press, 1998). Russell notes prior work in the field of econometrics, such as John Rust, "Do People Behave According to Bellman's Principle of Optimality?" Working Papers in Economics E-92-10 (The Hoover Institution, 1992).

[22] See, for example, Pieter Abbeel, Adam Coates, and Andrew Y. Ng, "Autonomous Helicopter Aerobatics through Apprenticeship Learning," *The International Journal of Robotics Research* 29 (13) (2010): 1608–1639.

[23] For taxi-driving, see Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey, "Maximum Entropy Inverse Reinforcement Learning," *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence* (Association for the Advancement of Artificial Intelligence, 2008). For sorting dishes, see Chelsea Finn, Sergey Levine, and Pieter Abbeel, "Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization," *Proceedings of the 33rd International Conference on Machine Learning* (International Conference on Machine Learning, 2016), 49–58.

[24] Felix Warneken and Michael Tomasello, "Altruistic Helping in Human Infants and Young Chimpanzees," *Science* 311 (5765) (2006). For more on infants' ability to understand others' goals, see Ashley J. Thomas, Christina M. Steele, Alison Gopnik, and Rebecca R. Saxe, "How Do Infants Experience Caregiving?" *Dædalus* 154 (1) (Winter 2025): 14–35, https://www.amacad.org/daedalus/how-do-infants-experience-caregiving.

[25] Simon Baron-Cohen, Alan M. Leslie, and Uta Frith, "Does the Autistic Child Have a 'Theory of Mind'?" *Cognition* 21 (1) (1985): 37–46.

[26] Michael Tomasello, Malinda Carpenter, Josep Call, et al., "Understanding and Sharing Intentions: The Origins of Cultural Cognition," *Behavioral and Brain Sciences* 28 (5) (2005): 675–691.

[27] Indeed, IRL was initially referred to as "apprenticeship learning" in the technical literature; see, for instance, Pieter Abbeel and Andrew Y. Ng, "Apprenticeship Learning Via Inverse Reinforcement Learning," *Proceedings of the Twenty-First International Conference on Machine Learning* (International Conference on Machine Learning, 2004).

[28] Eric Schwitzgebel, "Imagining Yourself in Another's Shoes versus Extending Your Concern: Empirical and Ethical Differences," *Dædalus* 154 (1) (Winter 2025): 134–149, https://www.amacad.org/daedalus/imagining-yourself-anothers-shoes-versus-extending-your-concern-empirical-ethical-differences.

[29] Note that both Thomas et al. and Schwitzgebel go on to highlight that care is more than the caregiver *simply* adopting the cared-for person's goals, preferences, and desires as their own–a point to which I will shortly return. Ibid.; and Thomas, Steele, Saxe, and Gopnik, "How Do Infants Experience Caregiving?"

[30] Dylan Hadfield-Menell, Stuart J. Russell, Pieter Abbeel, and Anca Dragan, "Cooperative Inverse Reinforcement Learning," *30th Conference on Neural Information Processing Systems* (Asia Pacific Neural Network Society, 2016), 29. Indexicality is a long-standing issue in the robotics literature; see, for example, David Chapman, *Vision, Instruction, and Action* (MIT Artificial Intelligence Laboratory, 1990); and Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza, "Power to the People: The Role of Humans in Interactive Machine Learning," *AI Magazine* 35 (4) (2014): 105–120.

[31] Paul F. Christiano, Jan Leike, Tom B. Brown, et al., "Deep Reinforcement Learning from Human Preferences," *Advances in Neural Information Processing Systems* 31 (2017).

[32] For summaries of documents, see Nisan Stiennon, Long Ouyang, Jeffrey Wu, et al., "Learning to Summarize with Human Feedback," *Advances in Neural Information Processing Systems* 33 (2020): 3008–3021. For preferences between binary responses, see Long Ouyang, Jeffrey Wu, Xu Jiang, et al., "Training Language Models to Follow Instructions with Human Feedback," *Advances in Neural Information Processing Systems* 35 (2022): 27730–27744.

[33] Not every LLM is necessarily fine-tuned with human feedback via RLHF, though as of 2024, the majority of the most widely used have been. Some LLMs are "base models" that have simply been trained to predict the next word (or linguistic "token") from a large body of text, with no further fine-tuning steps. Other LLMs, for instance the Claude series of models from Anthropic, are trained with a related but distinct process called Constitutional AI. See Yuntao Bai, Saurav Kadavath, Sandipan Kundu, et al., "Constitutional AI: Harmlessness from AI Feedback," arXiv, December 15, 2022.

[34] More explicitly democratic forms of AI alignment have been recently explored, including "collective constitutional AI" and "simulated mini-publics"; this represents an active area of research. See, respectively, Saffron Huang, Divya Siddarth, Liane Lovitt, et al., "Collective Constitutional AI: Aligning a Language Model with Public Input," *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (2024): 1395–1417; and Jan Leike, "A Proposal for Importing Society's Values," March 9, 2023, https://aligned.substack.com/p/a-proposal-for-importing-societys-values (accessed December 19, 2024).

[35] Bradbury, "The Veldt," 16.

[36] Ibid., 10.

[37] Ibid., 14.

[38] Ibid.

[39] Khalil Gibran, *The Prophet* (Alfred K. Knopf, 1923), 18.

[40] I'm grateful to Nim Tottenham for articulating this point during our workshop. See also Mary D. Salter Ainsworth, "Object Relations, Dependency, and Attachment: A Theoretical Review of the Infant-Mother Relationship," *Child Development* 40 (4) (1969): 969–1025.

[41] Arthur C. Clarke and Stanley Kubrick, *2001: A Space Odyssey*, Stanley Kubrick Productions, 1968.

[42] Bradbury, "The Veldt," 16.

[43] See Tristan Harris, "A Call to Minimize Distraction & Respect Users' Attention" (2013), http://minimizedistraction.com (accessed January 6, 2025).

[44] See, for example, Joel Lehman, "Machine Love," arXiv, February 22, 2023.

[45] Kent C. Berridge, Terry E. Robinson, and J. Wayne Aldridge, "Dissecting Components of Reward: 'Liking,' 'Wanting,' and Learning," *Current Opinions in Pharmacology* 9 (1) (2009): 65–73.

[46] Smith and Anderson, "Americans' Attitude toward Robot Caregivers"; Hurst, "Japan Lays Groundwork for Boom in Robot Carers"; Wright, "Inside Japan's Long Experiment in Automating Elder Care"; Wright, *Robots Won't Save Japan*; Schwartz, "Minnesota Nursing Homes Introduces [*sic*] Robot Caregivers"; Spichak, "Dementia Companion"; and Horowitz, "Who Will Take Care of Italy's Older People?"