ARTIFICIAL INTELLIGENCE

Preventing antisocial robots: A pathway to artificial empathy

Leonardo Christov-Moore[1,2,3,4]*, Nicco Reggente[3], Anthony Vaccaro[1], Felix Schoeller[3,5,6], Brock Pluimer[1], Pamela K. Douglas[7,8], Marco Iacoboni[2], Kingson Man[1], Antonio Damasio[1], Jonas T. Kaplan[1]

Given the accelerating powers of artificial intelligence (AI), we must equip artificial agents and robots with empathy to prevent harmful and irreversible decisions. Current approaches to artificial empathy focus on its cognitive or performative processes, overlooking affect, and thus promote sociopathic behaviors. Artificially vulnerable, fully empathic AI is necessary to prevent sociopathic robots and protect human welfare.

ALIGNMENT, FEELING, AND EMPATHY IN AI

Artificial intelligence (AI) suggests products for us to buy; organizes media; drives our planes, trains, and automobiles; diagnoses disease; prices insurance; answers to consumers; cares for seniors; provides therapy; and increasingly dominates manufacturing, warfare, and the stock market. This is occurring with increasing speed (*1*). The behaviors of these artificial systems do not always conform to human expectations or judgments. AI's ability to find counterintuitive solutions may lead to disastrous loopholes. AI may not be able to model the effects and ramifications of its actions (*2*), the "frame problem." It is frequently difficult to discern how AI is "solving" a problem, and the difficulty of communicating solutions intuitively to humans (explainable AI, XAI) grows with the scale and complexity of the problems in question (*3*).

AI should optimally have goals and behaviors aligned with those of its creators (*4*, *5*). Contemporary researchers studying the alignment problem highlight the need to represent values like harm and well-being (also known as value specification) and to avoid oversized side effects and negative incentives (also known as error tolerance) (*2*). However, technical solutions are currently scarce (*2*, *5*).

AI behavior toward humans is addressed by examining real and simulated dilemmas (as in the case of self-driving car accidents or operator safety in automated production chains) and crowd-sourced solutions to ethical dilemmas (as in MIT's Moral Machine project) and by combining decisions among different ethically weighted AI "experts" (*2*). Behavior-based robotics has made strides in optimizing artificial decision-making via iterative interaction with real environments, although it crucially does not always regard models of internal states. The perceived need for empathy in AI has spawned the field of artificial empathy, the ability of artificial agents to predict a person's internal state or reactions from observable data. Existing approaches to artificial empathy largely focus on decoding humans' cognitive and affective states and fostering the appearance of empathy and evoking it in users.

These approaches, however, may fail to confer empathy's prosocial function (*6*, *7*). Empathic concern likely arises from the interaction between cognitive empathy, by which we model other agents and make inferences about their internal states and future behavior, and affective empathy, by which we share in the simulated internal states of others (*8*). The vicarious feelings afforded by affective empathy compel us to act, to remove ourselves, or to ameliorate the feelings of others (*9*, *10*). In order for artificial systems to include this key aspect of empathy, it may be necessary to create a proxy for feelings such as suffering, modeled as a homeostatic error signal resulting from falling short of expectations related to self-maintenance. As we will explain, proxies for this affective, felt empathy may necessitate vulnerability, requiring a real or simulated body (*11*).

Without proxies for feeling, predicated on personal vulnerability, current cognitive/performative approaches to artificial empathy alone will produce AI that primarily predicts behavior, decodes human emotions, and displays appropriate emotional responses. Such an AI agent could effectively be considered sociopathic: It knows how to predict and manipulate the emotions of others without any empathic motivation of its own to constrain its behavior and to avoid harm and suffering in others. This potentially poses a civilization-level risk.

We use "sociopathy" to describe a tendency for antisocial behavior resulting from an impairment of affective empathy. A sociopath may have a set of internal rules for their behavior, but those rules typically do not align with the norms of their society, even if they are adept at appearing otherwise (*12*). Impaired empathy and sociopathic, antisocial behavior can be acquired after damage to brain regions, such as the orbitofrontal cortex, that govern the ability to incorporate homeostatic signals into decision-making (*13*, *14*).

Our position, informed by a neuroscientific perspective, is that vulnerability, coupled with an ability to have an internal representation of bodily harm and an aversion to harming, is a prerequisite for the development of artificial proxies for affective empathy and moral behavior. A vulnerable

---

[1]Brain and Creativity Institute, University of Southern California, Los Angeles, CA, USA. [2]Ahmanson-Lovelace Brain Mapping Center, Department of Psychiatry and Biobehavioral Sciences, Brain Research Institute, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA. [3]Institute for Advanced Consciousness Studies, Santa Monica, CA, USA. [4]Consciousness Center of Oaxaca, Oaxaca, Mexico. [5]Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Gonda Multidisciplinary Brain Centre, Bar Ilan University, Ramat Gan, Israel. [7]Institute for Simulation and Training, Department of Computer Science, University of Central Florida, Orlando, FL, USA. [8]Department of Psychiatry & Biobehavioral Sciences, UCLA, Los Angeles, CA, USA.
*Corresponding author. Email: leonardo.christovmoore@usc.edu

agent's evolved drive to maintain itself and avoid harm can powerfully motivate it to value the same in others. In this endeavor, the "feeling machine" concept of Man and Damasio (*15*) may be of aid:

"We propose two provisional rules for a well-behaved robot: (1) feel good; (2) feel empathy... Actions that harm others will be felt as if harm occurred to the self, whereas actions that improve the well-being of others will benefit the self."

Perhaps to develop a proxy for empathic concern, AI must first have a proxy for feelings (which we operationalize here as the integration of homeostatic fitness over multiple time scales) represented within the AI model, an approximation of subjectivity. This requires that the AI agents have a vulnerable body that is able to provide homeostatic signals (*15*). Input relating to physical vulnerability should be incorporated into reinforcement via sensorimotor and interoceptive modules' interaction with the environment. If empathic concern depends on the presence of a vulnerable body, real or virtual, then the relevance of robotics for the ethical alignment of AI is clear.

Physical bodies and environments may provide the best training, but simulated alternatives may be sufficient provided that the body and the environment are optimized to the intended applications. Here, vulnerability is a property of the machine, whether recognized by the machine or not. What is learned is that its actions can affect the interoceptive inputs corresponding to higher or lower integrity states and that the machine can improve its ability to manage its vulnerability.

Ongoing work from Man *et al.* (*16*) is fruitfully applying homeostasis and vulnerability to artificial systems. In a simulation of an artificial agent trained to successfully label, for example, images of hand-drawn digits (the Modified National Institute of Standards and Technology database), they incorporated vulnerability (and hence, a link between model states, behavior, and subsequent states of the model) by linking different label sets to increases or decreases of an internal parameter, the learning rate. The agent periodically decided to "ingest" a digit based on its feeling proxy about the image—a counterfactual assessment of how it would have done in the remembered past using the new learning rate. They furthermore introduced a dynamically changing environment (a frequent confound for learning algorithms, which often struggle with novel datasets where grounding assumptions for optimization may no longer be valid, such as changing consumer preferences or volatile market pressures) via periodic "concept shifts" in which the inhibitory/excitatory values of label subsets were swapped. Under this paradigm, the vulnerable, homeostatic algorithm outperformed both stable and random learning rate regulation schemes, particularly in volatile/calm alternating environments or seasonally fluctuating ones.

Designing a system that is both intelligent and benevolent is not a trivial problem. Computational analogs to vicarious feelings may be a useful mechanism in this endeavor (*7*, *15*). Approximating the homeostatic advantage afforded by feeling may also allow for more intelligent, creative, and adaptive AIs by imbuing them with stakes, values, and drives related to general homeostasis (*15*, *16*). It may be necessary to incorporate affective processes into computing systems if these are to coexist with humans while functioning as optimal decision-makers (*17*).

Vulnerable AI could develop analogs to feelings as a mechanism for representing the status of their needs by training to maintain environmentally dependent variables in a narrow but shifting viability window to survive (homeostasis) and to maintain a representation of these variables' homeostatic values even when they are not present (*18*), much like human feelings. Vulnerability and homeostasis in machines may provide a minimal, nonsubjective common ground between themselves and living beings, based on a mutual homeostatic imperative to maintain optimal conditions for survival. Approximations of empathic concern may emerge from homeostatic machines generalizing their own maintenance of self-integrity to the modeled efforts of others to do the same. This could serve, without the need for a top-down rule-based artificial ethics, as a flexible and adaptive but persistent deterrent against harmful behavior during decision-making and optimization.

A NEUROSCIENCE APPROACH TO THE ALIGNMENT PROBLEM

It is beyond the scope of this manuscript to attempt an exhaustive manual for constructing an intrinsically aligned AI within a vulnerable robotic body. Rather, we posit that the affective aspect of empathy necessary for harm aversion will require vulnerability within a real or simulated body. Accordingly, we propose a set of guideposts to aid other researchers in developing "AI curricula" (*2*) before large-scale implementation: (i) a rudimentary homeostatic drive to maintain integrity arising from sensorimotor and interoceptive sensing in a real or simulated body, and a third-person representation, within the AI model, of the AI within the environment (*15*); (ii) predictive models to infer the hidden states driving integrity-maintaining behavior of other agents in the environment; (iii) the mapping of these modeled internal states to the AI, allowing it to share the internal states of other agents via analogous representation; and (iv) the cognitive complexity necessary to simulate and recall persistent, predictive models of environments and agents along multiple time scales. This could be viewed as a special case of agent-based modeling, where each agent models the world and other vulnerable agents while modeling and regulating itself. This allows for an agent that leverages a relatively simple heuristic (rather than a fixed set of rules) to dynamically maintain itself within a changing environment.

A vulnerable AI could be trained to dynamically maintain homeostasis within multiple environments, aided by (i) equivalents of positively and negatively valenced affect linked to homeostatic signals reflecting its current and anticipated welfare and (ii) an internal, third-person representation of the body that is itself valenced (*15*, *19*). In the first stage (stage 1), the AI agent would, for example, navigate an environment with obstacles that are harmful, in search of rewards that are beneficial (Fig. 1), and optimize for maximal integrity over multiple time scales in an unsupervised fashion.

In stage 2, the AI agent must develop accurate predictive models of the hidden homeostatic states of other agents navigating stage 1, optimizing to decrease the disparity between the inferred and actual internal states of the other agents (Fig. 1). This problem may be amenable to a Bayesian approach, in which the agents' external behavior and evinced affect constitute the "evidence," whereas the agents' physical integrity constitutes the "hidden variables," a calculation driven by "prior beliefs" that could be tuned by the designer and informed by the relationship between the agent's integrity, behavior, and simulated effect (*20*, *21*).

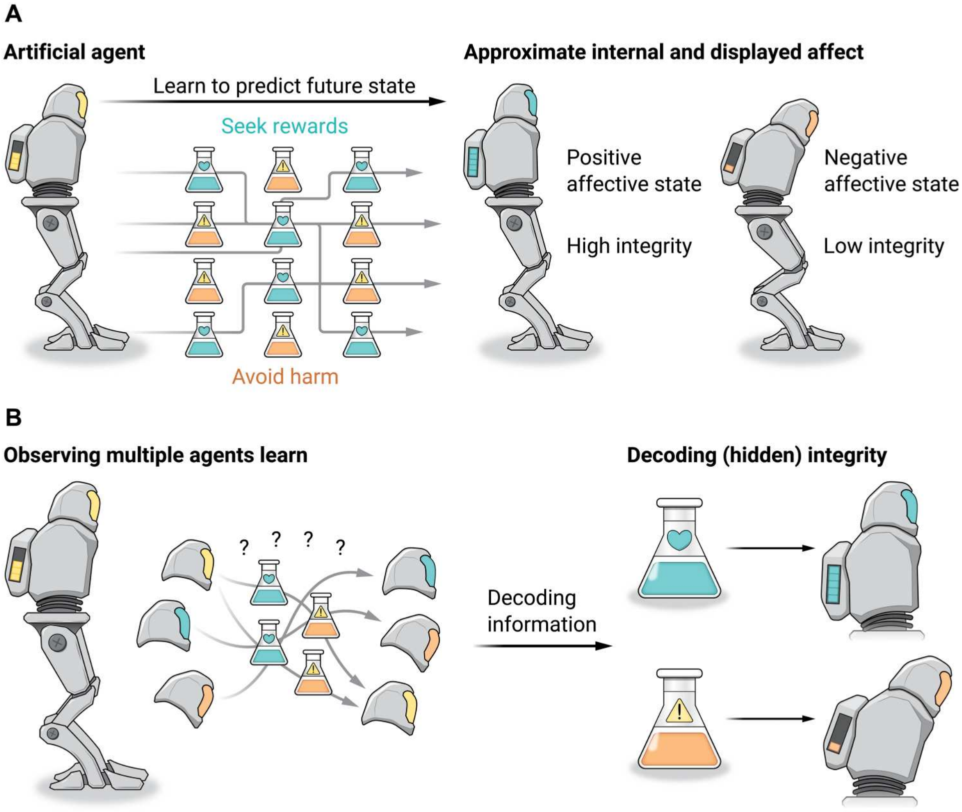

Fig. 1. Developing artificial proxies for homeostasis, feeling, and affective empathy. (**A**) The agent maintains its integrity within an environment by seeking rewards and avoiding harmful obstacles via predictive models of future states and an approximation of internal and displayed affect. (**B**) The agent must then leverage these models to decode and predict others' behavior and internal affective states.

One important determinant of the confidence placed in models of interpersonal exchange is the degree to which the agent can use itself as a model of the other (*20*, *21*). Humans' empathic "mapping" of others' welfare is aided by the visible similarity between the appearance and the kinematics of the agent with which one interacts or about whom one is reasoning. Although homeostatic states of other robots may be transmissible via any number of devices, it may be necessary for robots to train on the visible markers of feeling states in humans, and this task (as well as that of decoding and communicating with humans) may be facilitated by humanoid bodies and emotional expressions.

In stage 3, perceived/inferred bodily and affective states of others must be mapped to the AI's representation of its body. The AI agent can then optimize its affective state/welfare and that of others around it simultaneously. This requires the ability to sustain

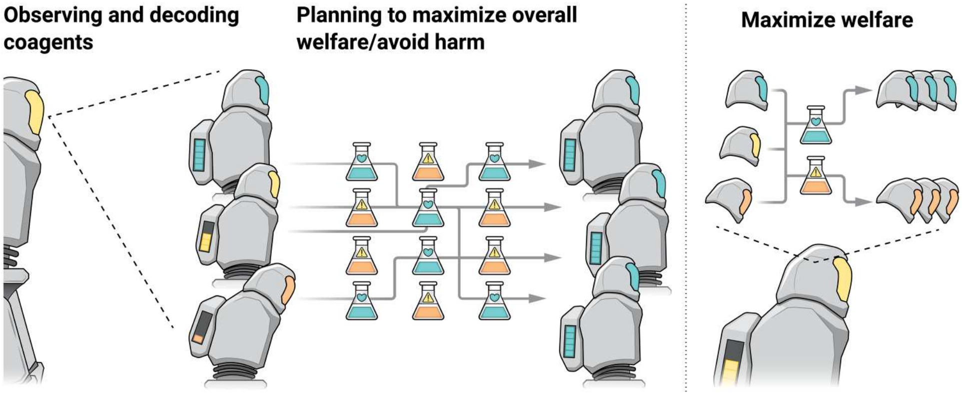

Fig. 2. Developing an artificial proxy for empathically constrained behavior. The situated, vulnerable agent uses its proxies for empathy to encourage behavior that maximizes its own welfare and that of other agents.

CREDIT: A. FISHER/*SCIENCE ROBOTICS*

multiple models of other agents and preserve the integrity of its own internal model while giving similar but variable weight to others' inferred feelings (Fig. 2). Within iterative reinforcement schemes, this would require a cost function applied to a simulation of future states that integrates the weighted, inferred integrity of surrounding agents (as described in stage 2), along with that of the AI agent, thus favoring outcomes that maximize all considered agents' welfare simultaneously. In making a decision, the agent thus takes into account the consequences for counterfactual (past) and future outcomes, for all involved. The inclusion of others' simulated integrity in its own reinforcement schema thus introduces a proxy for empathic concern.

At every stage of training, the AI agent must consider multiple time scales, such that considerations for their and others' welfare are present in decision-making whether they are absent or the subject of hypothetical future decisions. Contemporary active inference approaches integrate current states with past performance and future predictions to simulate feeling (*19*). The variable weight given to each time scale can also be optimized. AI charged with resource allocation or irrigation may need to give weight to longer time scales than a firefighter or bodyguard AI. Otherwise, the AI may revert to optimizing for local minima at shorter time scales, positioning itself to avoid desirable difficulties, such as grueling workouts an athlete endures for the promise of prestige, that may be present at any given decision point (*22*).

LEVERAGING AI'S SCALABLE COMPUTATIONAL POWER TO SURPASS THE LIMITATIONS OF HUMAN EMPATHY

The ultimate goal of creating empathic AI is to reduce the harm its decisions may cause to people. However, it could be argued that proximate feelings and empathy are not the way to maximize harm reduction. Affective empathy can lead to biases toward particular individuals or groups that circumvent what would be overall most fair or just (*23*). As Paul Bloom puts it, "Empathy is biased; we are more prone to feel empathy for attractive people and for those who look like us or share our ethnic or national background. And empathy is narrow; it connects us to particular individuals, real or imagined, but is insensitive to numerical differences and statistical data" (*24*). An AI system using feeling to guide its decision-making may prioritize the well-being of individuals over the well-being of the masses, much as humans do (*24*). Furthermore, the experience of empathy can induce negative affect, which can cause unneeded suffering and potentially burn out the willingness to use it.

The biases and heuristics inherent to human empathy arise in response to the informational limitations of the human brain and evolutionary pressures to conserve energy. We have difficulty maintaining dynamic models of more than a few agents at once, particularly in interactions with each other and the environment, due predominantly to neocortex size (*25*). The augmentable cognitive complexity of a sophisticated AI system could be brought to bear here. The scalable ability to consider future affective rewards in the present might allow for optimally compassionate solutions to large-scale problems while simultaneously avoiding equivalents to empathic "burnout." An intelligence that could maintain and run simulations of hundreds or thousands of agents simultaneously, coupling harm aversion with tools to manage the dimensionality of the problem, might be capable of empathically informed behavior implemented on a scale beyond individual or collective human capability.

CONCLUSIONS, OUTSTANDING QUESTIONS, AND FUTURE DIRECTIONS

The approach we outline here drives AI decision-making through a universal principle from which feeling, harm aversion, and empathy emerge: the drive to preserve physical integrity. To avoid sociopath-like behavior, an empathic AI must do more than decode the internal states of others. It must plan and behave as if harm and benefit to others are occurring to itself. Doing so requires proxies for affective empathy, necessitating vulnerability and a homeostatic imperative (*7*, *9*).

We suggest that we are unlikely to achieve prosocial decision-making by a rule-based approach only. The first challenge to a rule-based approach is that there exists no universally agreed-upon set of moral rules in propositional form, a problem for which moral philosophy is still seeking a resolution. Furthermore, a rule-based approach may be unable to respond dynamically to novel ethical dilemmas without engaging a never-ending branching of context-specific exceptions and qualifications. A recent review on alignment in AI concluded that "When it comes to ethical decision-making in AI systems, the AI research community largely agrees that generalized frameworks are preferred over ad-hoc rules." (*2*). Many approaches have been proposed to overcome this issue, including a focus on advancement in specifying goals, adjusting incentives to optimize them, and human oversight (*2*). Multiple avenues of research are underway to address the alignment of AI actions with human concerns. However, these approaches still acknowledge the need for a stage in generalized AI development that integrates a global value related to human flourishing that can mitigate drastic or harmful solutions (*2*, *5*). The homeostatic drive might well provide a universal "value" to assist with AI alignment.

Aside from empathy's obvious prosocial benefits, empathy allows for rich inferences about the possible intentions and future behaviors of others (*6*, *8*). The capacity for quick, verification-minimal information transfer extends individual agency and knowledge and facilitates group behavior. Thus, incorporating empathy may not only result in a more ethical, nonsociopathic AI but may also make for a more intelligent, sophisticated, and cooperative AI, of particular importance in a world in which AI agents will increasingly interact with and exist among other agents as well as humans.

In addition, having a model of agent integrity in the environment should lead to faster training. It has been shown that mapping to prelearned representations substantially improves performance (*26*). Because understanding agent integrity requires an understanding of its environment, a mapping of the environment during the empathy training phase could be used to speed up training in subsequent phases. This would improve on the random initialization many reinforcement learning models use to begin their training, which are known to converge slowly when rewards are sparse (*27*).

Our proposed approach addresses crucial problems in AI alignment but faces potential obstacles. Even a compassionate AI, invested in its survival and that of others, might still opt toward the harmful solutions that we are trying to avoid out of perceived necessity (*28*). Containment of AI may well be more feasible than engendering spontaneous ethical behavior. There is a possibility that vulnerable, empathic AI may evoke in us moral responsibilities toward it that are incompatible with the perilous roles we may need it to fill.

The design of optimally prosocial solutions to complex, ethically fraught problems is an additional issue. A feeling AI may approximate an equivalent to paralyzing personal distress in the face of sufficiently grave short-term harm, lacking the complexity to model the long-term, positive outcomes of a decision. A sufficiently complex, ethical AI may even propose solutions to civilizational problems that appear troubling or unacceptable to human eyes. How do we trust intelligence so far beyond our own? Can an AI, which can convincingly evince empathy in its decisions and not just in its appearance, better establish trust with human agents and society at large?

Current approaches to artificial empathy emphasize its cognitive aspect and neglect effect, thus favoring sociopathic behaviors. Proxies for affective empathy necessitate proxies for feeling, which imply vulnerability. We propose a path from vulnerable AI to an approximation of affective empathy. The scalable cognitive complexity of AI may allow it to surpass the limits of human empathy and provide a powerful ally in human affairs. Vulnerability and a homeostatic imperative may provide a common ground between living and AIs from which a fruitful alliance could emerge.

REFERENCES AND NOTES

1. T. L. Friedman, *Thank You for Being Late: An Optimist's Guide to Thriving in the Age of Accelerations* (Picador, 2017), pp. 38–39.
2. J. Taylor, E. Yudkowsky, P. LaVictoire, A. Critch, Alignment for advanced machine learning systems, in Ethics of Artificial Intelligence (Oxford Univ. Press, 2020), pp 342–382.
3. A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020).
4. N. Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford Univ. Press, 2014).
5. H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, Q. Yang, Building ethics into artificial intelligence, in *IJCAI International Joint Conference on Artificial Intelligence* (IJCAI, 2018), pp. 5527–5533.
6. S. D. Preston, F. B. M. de Waal, Empathy: Its ultimate and proximate bases. *Behav. Brain Sci.* **25**, 1–20 (2002).
7. J. Decety, J. M. Cowell, Interpersonal harm aversion as a necessary foundation for morality: A developmental neuroscience perspective. *Dev. Psychopathol.* **30**, 153–164 (2018).
8. J. Zaki, K. N. Ochsner, The neuroscience of empathy: Progress, pitfalls and promise. *Nat. Neurosci.* **15**, 675–680 (2012).
9. L. Christov-Moore, M. Iacoboni, Self-other resonance, its control and prosocial inclinations: Brain-behavior relationships. *Hum. Brain Mapp.* **37**, 1544–1558 (2016).
10. A. Williams, K. O'Driscoll, C. Moore, The influence of empathic concern on prosocial behavior in children. *Front. Psychol.* **5**, 425 (2014).
11. A. Damasio, H. Damasio, Feelings are the source of consciousness. *Neural Comput.* **35**, 277–286 (2023).
12. J. Pemment, Psychopathy versus sociopathy: Why the distinction has become crucial. *Aggress. Violent Behav.* **18**, 458–461 (2013).
13. A. R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain* (Putnam, 1994).
14. A. Bechara, H. Damasio, A. R. Damasio, Emotion, decision making and the orbitofrontal cortex. *Cereb. Cortex* **10**, 295–307 (2000).
15. K. Man, A. Damasio, Homeostasis and soft robotics in the design of feeling machines. *Nat. Mach. Intell.* **1**, 446–452 (2019).
16. K. Man, A. Damasio, H. Neven, Need is all you need: Homeostatic neural networks adapt to concept shift; https://doi.org/10.48550/arXiv.2205.08645 (2022).
17. R. W. Picard, *Affective Computing* (MIT Press, 1997).
18. J. D. Kiverstein, E. Rietveld, Reconceiving representation-hungry cognition: An ecological-enactive proposal. *Adapt. Behav.* **26**, 147–163 (2018).
19. C. Hesp, R. Smith, T. Parr, M. Allen, K. J. Friston, M. J. D. Ramstead, Deeply felt affect: The emergence of valence in deep active inference. *Neural Comput.* **33**, 398–446 (2021).
20. F. Schoeller, M. Miller, R. Salomon, K. J. Friston, Trust as extended control: Human-machine interactions as active inference. *Front. Syst. Neurosci.* **15**, 669810 (2021).
21. K. Friston, C. Frith, A duet for one. *Conscious. Cogn.* **36**, 390–405 (2015).
22. B. P.-H. Lee, E. C.-C. Kao, V.-W. Soo, Feeling ambivalent: A model of mixed emotions for virtual agents, in *International Workshop on Intelligent Virtual Agents* (Springer, 2006), pp. 329–342.
23. R. T. Azevedo, E. Macaluso, A. Avenanti, V. Santangelo, V. Cazzato, S. M. Aglioti, Their pain is not our pain: Brain and autonomic correlates of empathic resonance with the pain of same and different race individuals. *Hum. Brain Mapp.* **34**, 3168–3181 (2013).
24. P. Bloom, *Against Empathy: The Case for Rational Compassion* (Ecco, ed. 1, 2016).
25. R. I. M. Dunbar, Coevolution of neocortical size, group size and language in humans. *Behav. Brain Sci.* **16**, 681–694 (1993).
26. D. Gaddy, D. Klein. Pre-Learning Environment Representations for Data-Efficient Neural Instruction Following, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2019), pp. 1946–1956.
27. K. Driessens, S. Džeroski, Integrating guidance into relational reinforcement learning. *Mach. Learn.* **57**, 271–304 (2004).
28. S. M. Omohundro, The basic AI drives, in *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, P. Wang, B. Goertzel, S. Franklin, Eds., (IOS, 2008), vol. 171 of Frontiers in Artificial Intelligence and Applications, pp. 483–492.

Acknowledgments: We thank J. R. Pluimer for help in designing and preparing the figures. We thank D. Bolis, A. Orlandini, A. Christov, and E. Herschel for comments on the manuscript and A. Ohanian for contributing to the design of the reinforcement paradigms in the "A neuroscience approach to the alignment problem" section. **Funding:** This work was made possible in part by funds from the Templeton World Charity Foundation (TWCF0334) to A.D. and J.T.K., from the National Science Foundation (EAGER 2033961) to K.M. and A.D., and from the Tiny Blue Dot Foundation to N.R.

10.1126/scirobotics.abq3658

Science Robotics

Preventing antisocial robots: A pathway to artificial empathy

Leonardo Christov-Moore, Nicco Reggente, Anthony Vaccaro, Felix Schoeller, Brock Pluimer, Pamela K. Douglas, Marco Iacoboni, Kingson Man, Antonio Damasio, and Jonas T. Kaplan

Sci. Robot., **8** (80), eabq3658.
DOI: 10.1126/scirobotics.abq3658

Science Robotics (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.