Comment

SEAN GALLUP/GETTY

A huddle at the 2017 United Nations Climate Change Conference, where attendees cooperated on mutually beneficial joint actions on climate.

Cooperative AI: machines must learn to find common ground

Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson & Thore Graepel

To help humanity solve fundamental problems of cooperation, scientists need to reconceive artificial intelligence as deeply social.

Artificial-intelligence assistants and recommendation algorithms interact with billions of people every day, influencing lives in myriad ways, yet they still have little understanding of humans. Self-driving vehicles controlled by artificial intelligence (AI) are gaining mastery of their interactions with the natural world, but they are still novices when it comes to coordinating with other cars and pedestrians or collaborating with their human operators.

The state of AI applications reflects that of the research field. It has long been steeped in a kind of methodological individualism. As is evident from introductory textbooks, the canonical AI problem is that of a solitary machine confronting a non-social environment. Historically, this was a sensible starting point. An AI agent — much like an infant — must first master a basic understanding of

its environment and how to interact with it.

Even in work involving multiple AI agents, the field has not yet tackled the hard problems of cooperation. Most headline results have come from two-player zero-sum games, such as backgammon, chess[1], Go[2] and poker[3]. Gains in these competitive examples can be made only at the expense of others. Although such settings of pure conflict are vanishingly rare in the real world, they make appealing research projects. They are culturally cherished, relatively easy to benchmark (by asking whether the AI can beat the opponent), have natural curricula (because students train against peers of their own skill level) and have simpler solutions than semi-cooperative games do.

AI needs social understanding and cooperative intelligence to integrate well into society. The coming years might give rise to diverse ecologies of AI systems that interact in rapid and complex ways with each other and with humans: on pavements and roads, in consumer and financial markets, in e-mail communication and social media, in cybersecurity and physical security. Autonomous vehicles or smart cities that do not engage well with humans will fail to deliver their benefits, and might even disrupt stable human relationships.

We need to build a science of cooperative AI. As researchers in the field and its governance, we argue that it is time to prioritize the development of cooperative intelligence that has the ability to promote mutually beneficial joint action, even when incentives are not fully aligned. Just as psychologists studying humans have found that the infant brain does not develop fully without social interaction, progress towards socially valuable AI will be stunted unless we put the problem of cooperation at the centre of our research.

Cooperative intelligence is unlikely to emerge as a by-product of research on other kinds of AI. We need more work on cooperative games and complex social spaces, on understanding norms and behaviours, and on social tools and infrastructure that promote cooperation. The AI community should learn more from, and contribute to, other fields that work on cooperation.

From autonomy to cooperation

Parents encourage their children to grow beyond their dependencies and become autonomous. But autonomy is rarely regarded as the sole goal for humans. Rather, we are generally most productive when we work cooperatively as part of broader society. Similarly, certain kinds of autonomy in AI systems are useful precisely because they enable the system to contribute effectively to broader cooperative efforts. Most of the value from self-driving vehicles will come not from driving on empty roads, but from vehicles coordinating smoothly with the flow of pedestrians, cyclists and cars driven by humans. Thus, cooperative intelligence is not an alternative to autonomous intelligence, but goes beyond it.

Four elements of cooperative intelligence

In most settings, people's incentives are not fully aligned. Nevertheless, they can often cooperate, taking joint action to achieve mutually beneficial outcomes. One example is countries agreeing and enforcing carbon cuts to tackle climate change. The cooperative intelligence needed to achieve this has four parts:

Understanding. The ability to take into account the consequences of actions, to predict another's behaviour, and the implications of another's beliefs and preferences.

Communication. The ability to explicitly and credibly share information with others relevant to understanding behaviour, intentions and preferences.

Commitment. The ability to make credible promises when needed for cooperation.

Norms and institutions. Social infrastructure — such as shared beliefs or rules — that reinforces understanding, communication and commitment.

AI research on cooperation will need to bring together many clusters of work. A first cluster consists of AI–AI cooperation, tackling ever more difficult, rich and realistic settings (see 'Four elements of cooperative intelligence'). A second is AI–human cooperation, for which we will need to advance natural-language understanding, enable machines to learn about people's preferences, and make machine reasoning more accessible to humans. A third cluster is work on tools for improving (and not harming) human–human cooperation, such as ways of making the algorithms that govern social media better at promoting healthy online communities.

AI–AI cooperation

Multi-agent AI research has seen most success in two-player zero-sum settings, from the superhuman performance of IBM's chess-playing computer Deep Blue to the powerful demonstration of deep reinforcement learning by the program AlphaGo. However, few interactions in the real world are characterized by pure conflict — when there is no possibility of bargains, negotiations or threats. So, improving skill at inherently rivalrous games is unlikely to be the most promising way for AI to produce social value.

Games of pure common interest are a step towards developing cooperative agents. The cooperative card game Hanabi[4] requires players to communicate private information and intentions under strong constraints about what can be said and when. Team games such as robot soccer[5] need players on a team to work as one, jointly planning their moves and passing the ball. In these examples, all agents on a team share the same goals. Mastering these games requires many skills essential to cooperation. Research avenues include building AIs that can understand what teammates are thinking and planning; communicate plans; and even cooperate with different kinds of teammate who might think differently and react more slowly (known as ad hoc teamwork).

Yet because these situations are restricted to a perfect harmony of interests, they represent the easy case for cooperation. Real-world relationships almost always involve a mixture of common and conflicting interests. This tension gives rise to the rich texture of human cooperation problems, including bargaining, trust and mistrust, deception and credible communication, commitment problems and assurances, politics and coalitions, and norms and institutions. AI agents will need to learn how to manage these harder cooperation problems, as humans do.

An example is the board game Diplomacy, in which players negotiate non-binding alliances with others. To succeed, AI agents will need to understand each other well enough to recognize when their interests are aligned with those of other players. They will have to develop a common vocabulary to communicate their intentions. They will benefit from being able to communicate credibly, despite possible incentives to lie. They must overcome mutual fears of betrayal, so as to agree on and execute jointly beneficial plans. They might even learn to establish norms relating to the adherence of agreements. To enable progress in these cooperative skills, researchers have devised variants of Diplomacy that modify the difficulty of these challenges, such as introducing an agreed simple vocabulary or permitting binding commitments.

Human–AI cooperation

AI is increasingly present, underlying everything from dynamic pricing strategies to loans and prison-sentencing decisions. Collaborative industrial robots work on factory floors alongside labourers[6], care robots help human health workers and personal AI assistants (such as Amazon's Alexa, Apple's Siri and Google Assistant) help us with scheduling, albeit in an elementary way.

DPA PICTURE ALLIANCE/ALAMY

Computer scientists in Leipzig, Germany, prepare their robot soccer team for a test game.

The design of agents that will act in accordance with human intentions, preferences and values — known as AI alignment[7] — is a crucial part of cooperative AI. But it is only a part, because the relationship between a single human (acting as the principal) and a single machine (acting as the agent) isn't always clear. Real-world cooperation problems often involve multiple stakeholders, some conflicting interests and integration with our institutional and normative infrastructure.

A particular challenge facing researchers working on human–AI cooperation is that it involves, well, humans. Today, many deployed machine-learning models are trained either on massive data sets or in simulated environments that can generate years of experience in seconds. For example, the program AlphaZero learnt to play chess by playing 44 million games against itself over 9 hours. By contrast, humans produce data slowly, and require researchers to consider compensation, ethics and privacy. There might be ways to use fewer human participants, such as extensively training AIs in simulation and then fine-tuning them in the real world.

Many AI practitioners have dreamt of building autonomous and human-like intelligence. They envisioned systems that could replace human labour, acting with the speed and resilience of machines, and scaling up rapidly given increases in computing power, algorithmic efficiency and capital. However, unlike systems that have tight integration with human workers, autonomous systems might pose greater safety risks. Human-like AI might be more likely to displace labour.

Instead, we could develop AI assistants that complement human intelligence and depend on us for tasks in which humans have a comparative advantage. As Stanford University radiologist Curtis Langlotz put it: "AI won't replace radiologists, but radiologists who use AI will replace radiologists who don't."

> **"To succeed, cooperative AI must connect with the broader science of cooperation."**

Progress will require advances in understanding human language, gestures and activities, and ad hoc teamwork, in addition to preference learning by machines, safety, interpretability by humans[8], and understanding of norms. Research will need to approach increasingly rich and realistic environments. Instead of benchmarking progress mainly by whether autonomous machines can outperform autonomous humans on a task, researchers should also assess the performance of human–machine teams.

AI for human collaboration

Humans confront ubiquitous cooperation problems as commuters, neighbours, co-workers and citizens. The global scientific community, for example, could benefit from better tools for identifying relevant work and promising collaborations. Technology is crucial, mediating our ability to find and process information, communicate and self-organize. Digital systems and AI can expand this toolkit.

Some AI tools, such as machine language translation, seem strongly disposed towards promoting cooperation. Today, 2 people who speak any of more than 100 languages can communicate with the aid of a smartphone and a translation app.

Digital platforms such as Wikipedia, Reddit and Twitter provide tools to combine user-provided content. AI advances could improve this community infrastructure, for example, by routing relevant information to contributors more efficiently to enhance collaborative editing. Other advances could improve user rating and reputation systems through better modelling and by accounting for the rater's repute or relevance, as well as by enabling recommendation algorithms that more

People who speak different languages can communicate using an AI-based translation device.

AKIO KON/BLOOMBERG VIA GETTY

intelligently promote a community's values.

Building healthy online communities is challenging; just as social media can connect us, so too can it polarize, stress, misinform, distract and addict us[9]. Researchers and developers need to find better ways to name and measure desirable properties and build algorithms that encourage them. Platforms for political deliberation can be designed to promote empathy about different viewpoints and cultivate community consensus. Methods for achieving this include language comprehension that links to structured databases of knowledge, or clustering algorithms to identify related perspectives.

Next steps

To succeed, cooperative AI must connect with the broader science of cooperation, which spans the social, behavioural and natural sciences. AI research will need to converse with multiple fields. These include psychology, to understand human cognition; law and policy, to understand institutions; history, sociology and anthropology, to understand culture; and political science and economics, to understand problems of information, commitment and social choice. Adjacent research areas are developing AI with socially desirable properties, such as alignment, interpretability and fairness[10,11]. Each of these addresses a distinct, but complementary, set of challenges.

The need for interdisciplinarity is exemplified by a landmark work: Robert Axelrod's *The Evolution of Cooperation*, published in 1984 (ref .12). Axelrod, a political scientist, brought together game theorists, mathematicians, economists, biologists and psychologists in a tournament to help devise the best algorithms for the iterated Prisoner's Dilemma, the canonical example of why two rational people might not cooperate. The winning solution that cooperated most successfully, called Tit for Tat, was devised by Anatol Rapoport, a US scholar with a background spanning mathematics, biology, network science and peace studies.

Axelrod's tournament offered another lesson. It gave researchers a benchmark for success in the design of cooperative algorithms, just as ImageNet[13] did for computer vision by collecting and labelling millions of photos. Cooperative AI research will similarly gain momentum if investigators can devise, agree on and adopt benchmarks that cover a diverse set of challenges: playing cooperative board games, integrating into massive multiplayer video games, navigating simplified environments that require machine–human interaction, and anticipating tasks as a personal assistant might. Similar to the state-of-the-art in language modelling, considerable effort and creativity will be needed to make sure these benchmarks remain sufficiently rich and ambitious, and do not have socially harmful blind spots.

> **"Promoting research into cooperative AI will require social interventions."**

The most important challenges of cooperation might be the most difficult to benchmark; they involve creatively stepping out of our habitual roles to change the 'game' itself. Indeed, if we are to take the social nature of intelligence seriously, we need to move from individual objectives to the shared, poorly defined ways humans solve social problems: creating language, norms and institutions.

Science is a social enterprise, so promoting research into cooperative AI will require social interventions. A recent milestone was a December 2020 workshop on cooperative AI at the leading machine-learning conference NeurIPS. It involved speakers from a diverse array of disciplines, and resulted in a review of Open Problems in Cooperative AI[14].

We and others are establishing a Cooperative AI Foundation to support this nascent field (www.cooperativeai.org), backed by a large philanthropic commitment. The foundation's mission will be to catalyse advances in cooperative intelligence to benefit all of humanity, including efforts to fund fellowships, organize conferences, support benchmarks and environments, and award prizes.

The crucial crises confronting humanity are challenges of cooperation: the need for collective action on climate change, on political polarization, on misinformation, on global public health or on other common goods, such as water, soil and clean air. As the potential of AI continues to scale up, a nudge in the direction of cooperative AI today could enable us to achieve much-needed global cooperation in the future.

The authors

Allan Dafoe is associate professor and director at the Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, UK. **Yoram Bachrach** is a research scientist at DeepMind, London. **Gillian Hadfield** is director at the Schwartz Reisman Institute for Technology and Society, University of Toronto, Canada; and senior policy adviser at OpenAI, San Francisco, California, USA. **Eric Horvitz** is chief scientific officer at Microsoft, Redmond, Washington, USA. **Kate Larson** is professor of computer science at the University of Waterloo, Canada, and a research scientist at DeepMind, Montreal, Canada. **Thore Graepel** is research lead at DeepMind, London; and professor of computer science at University College London.
e-mails: allan.dafoe@governance.ai; thore@google.com

1. Campbell, M., Hoane Jr, A. J. & Hsu, F.-H. *Artif. Intell.* **134**, 57–83 (2002).
2. Silver, D. *et al. Nature* **529**, 484–489 (2016).
3. Moravčík, M. *et al. Science* **356**, 508–513 (2017).
4. Bard, N. *et al. Artif. Intell.* **280**, 103216 (2020).
5. Kitano, H. *et al. AI Magazine* **18**, 73 (1997).
6. Villani, V., Pini, F., Leali, F. & Secchi, C. *Mechatronics* **55**, 248–266 (2018).
7. Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control* (Penguin, 2019).
8. Adadi, A. & Berrada, M. *IEEE Access* **6**, 52138–52160 (2018).
9. Allcott, H., Braghieri, L., Eichmeyer, S. & Gentzkow, M. *Am. Econ. Rev.* **110**, 629–676 (2020).
10. Barocas, S., Hardt, M. & Narayanan, A. *Fairness and Machine Learning* (Fairmlbook.org, 2019).
11. Buolamwini, J. & Gebru, T. *Proc. Mach. Learn. Res.* **81**, 77–91 (2018).
12. Axelrod, R. *The Evolution of Cooperation* (Basic, 1984).
13. Deng, J. *et al. IEEE Conf. Comput. Vis. Pattern Recognit.* **2009**, 248–255 (2009).
14. Dafoe, A. *et al.* Preprint at https://arxiv.org/abs/2012.08630 (2020).