Social impact and governance of AI and neurotechnologies

Neural Networks 152 (2022) 542–554

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2021 Special Issue on AI and Brain Science: Brain-inspired AI

Social impact and governance of AI and neurotechnologies

Kenji Doya [a,*], Arisa Ema [b], Hiroaki Kitano [a,c], Masamichi Sakagami [d], Stuart Russell [e]

[a] Neural Computation Unit, Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna Village, Okinawa, 904-0495, Japan
[b] Institute for Future Initiatives, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
[c] Sony Computer Science Laboratories, Inc. 3-14-13, Higashigotanda, Shinagawa-ku, Tokyo, 141-0022, Japan
[d] Brain Science Research Center, Tamagawa University, 6-1-1 Tamagawa Gakuen, Machida City, Tokyo, 194-8610, Japan
[e] Computer Science Division, University of California Berkeley, CA 94720-1776, USA

ARTICLE INFO

Article history:
Available online 21 May 2022

Keywords:
Artificial intelligence
Neurotechnology
AI scientist
Human compatible AI
Ethics
Governance

ABSTRACT

Advances in artificial intelligence (AI) and brain science are going to have a huge impact on society. While technologies based on those advances can provide enormous social benefits, adoption of new technologies poses various risks. This article first reviews the co-evolution of AI and brain science and the benefits of brain-inspired AI in sustainability, healthcare, and scientific discoveries. We then consider possible risks from those technologies, including intentional abuse, autonomous weapons, cognitive enhancement by brain–computer interfaces, insidious effects of social media, inequity, and enfeeblement. We also discuss practical ways to bring ethical principles into practice. One proposal is to stop giving explicit goals to AI agents and to enable them to keep learning human preferences. Another is to learn from democratic mechanisms that evolved in human society to avoid over-consolidation of power. Finally, we emphasize the importance of open discussions not only by experts, but also including a diverse array of lay opinions.

1. Introduction

Artificial intelligence (AI) has made remarkable progress in the last several years. Computers can now see, listen, and drive a car, in some cases as well as humans, or even better. Machine translation is no longer an embarrassment. Much of today's AI is based on deep learning (DL), a brain-inspired machine learning framework (Goodfellow et al., 2016; Sejnowski, 2018). Together with other brain-inspired and statistical approaches, AI is expected to further benefit human life and society in areas such as information services, manufacturing, mobility, environment, healthcare, and science. However, rapid advances also pose large risks. Here we consider expected benefits and potential risks of AI and neurotechnologies and how these technologies can be managed so as to minimize undesirable outcomes.

This article is based on presentations and discussions at the International Symposium on AI and Brain Science held online in October 2020 (http://www.brain-ai.jp/symposium2020/). In the following sections, we first review how AI and brain science have co-evolved and what else can be learned from the brain to guide future progress of AI. We then review how progress in AI and neurotechnologies can deliver helpful innovations, with a focus on the potential of AI to promote scientific discoveries. We then consider the risks associated with AI and neurotechnologies, including intentional misuse, unintended side effects, and loss of control. We then consider ethics and governance of AI, with two specific viewpoints. For the issue of goal misspecification by AI systems, a theoretical framework of assistance games is proposed, in which AI agents continually infer human preferences, rather than pursuing a specific goal. Another view proposes, by learning from the history of human society, to implement democratic peer reviewing among open-source, explainable AI agents, to avoid catastrophes by over-concentration of power.

2. Co-evolution of AI and neuroscience

There are two opposing views of the relationship between AI and the brain. One holds that to make intelligent machines with electronics, we do not need to be concerned with biological constraints. However, the other maintains that since there is already a superb implementation of intelligence in the brain, it makes more sense to reverse engineer that. In fact, there are many intermediate views and historically, dominant views have oscillated between the two extremes.

For example, in visual pattern recognition, the discovery by Hubel and Wiesel (Hubel & Wiesel, 1959) that the primary visual cortex is composed of neurons that respond to different local

* Corresponding author.
E-mail addresses: doya@oist.jp (K. Doya), ema@ifi.u-tokyo.ac.jp (A. Ema), kitano@sbi.jp (H. Kitano), sakagami@lab.tamagawa.ac.jp (M. Sakagami), russell@berkeley.edu (S. Russell).

https://doi.org/10.1016/j.neunet.2022.05.012

features, promoted development of the Perceptron (Rosenblatt, 1957), which adaptively combines those features for pattern classification. While the limitation of simple Perceptrons (Minsky & Papert, 1969) brought a winter to neural network research, researchers developed learning algorithms for multi-layer neural networks to allow learning of simple to complex features (Amari, 1967; Carpenter & Grossberg, 1987; Fukushima, 1980; Rumelhart et al., 1986; Werbos, 1974). Neuroscientists also discovered high-level feature representations, like face cells in the higher visual cortex (Bruce et al., 1981) and place cells in the hippocampus (O'Keefe, 1976). Most recently, with the availability of big data from the internet and high-performance parallel computers, deep neural networks began to achieve very high performance (Krizhevsky et al., 2012).

In regard to action learning, experimental psychology and neuroscience revealed that prediction of forthcoming reward, based on sensory cues or actions, is critical for action learning (Rescorla & Wagner, 1972). Inspired by such notions, Barto and colleagues developed the computational framework of reinforcement learning, in which the reward prediction error is used as the key learning signal (Barto et al., 1983; Sutton & Barto, 2018). The subsequent discovery by Schultz that dopaminergic neurons signal reward prediction error (Schultz et al., 1993) led to development of models of reinforcement learning in brain circuits (Barto, 1995; Schultz et al., 1997), which further promoted neuroscience studies of reinforcement learning in the brain (O'Doherty et al., 2003; Samejima et al., 2005). While applications of reinforcement learning remained dormant due to its instability when combined with multi-layer neural networks (Boyan & Moore, 1995), discovery of a stable way to use deep neural networks for reinforcement learning, with insights from hippocampal episodic memory (Hassabis et al., 2017), brought flourishing of deep reinforcement learning (Mnih et al., 2015), which enables machines to defeat human champions in the game of Go (Silver et al., 2016).

2.1. What can we learn from the brain for advancement of AI

As mentioned above, deep learning for pattern recognition and reinforcement learning for action and planning are two prime examples of brain-inspired AI (Hassabis et al., 2017). What else can we learn from the brain to advance next-generation AI? There are at least three domains in which present AI cannot currently match the human brain and can therefore benefit from advances in brain science.

One domain is energy efficiency. Today's deeper learning depends on GPUs that consume large quantities of energy and produce a lot of heat. On the other hand, our brain is thought to use just about 20 watts of energy. Such energy efficiency presumably derives from distributed computation by dendrites, molecular machinery in neurons, and efficient communication via spikes. Neuromorphic engineering based on distributed analog computing and spike communication is an active research area (Roy et al., 2019).

Another important domain is data efficiency. Humans appear to be able to acquire new knowledge or skills with much less experience than today's deep neural networks require. There are several reasons for data efficient learning, such as mental simulation, modularity and compositionality, and meta-learning (Lake et al., 2017).

The third important domain is autonomy and sociality. Today's AI agents are designed by human engineers regarding what is to be achieved, but humans and animals are not designed or dictated to do something. They decide what to do and what to learn so as to function in the physical, biological, and social environments into which they are born. How we can create autonomous or creative AI agents like humans is a fascinating and controversial issue (Elfwing et al., 2011; Oudeyer, 2018; Schwartenbeck et al., 2019).

2.2. Toward data-efficient learning

Understanding the reasons for data-efficient learning by the brain and to achieve data-efficient learning by artificial agents are targets of active research. An important factor is the use of predictive models, which allow "mental simulation" in animals and humans (Hamrick, 2019). Mental simulation can be defined as the brain's process using action-dependent state transition models. It allows estimation of the present state based on past sequences of actions and observations, planning of future actions from the present state, or imagination of possible actions from arbitrary states. Mental simulation allows efficient adaptation when task requirements are changed. Mental simulation also enables understanding and explaining what has happened and why a certain action is preferred. The brain's mechanisms for mental simulation are being revealed by advanced neuroimaging and data analysis (Fermin et al., 2016; Funamizu et al., 2016; Soto et al., 2020). Learning and use of action-dependent state transition models are now becoming popular in deep learning architectures to allow flexible adaptation (Eslami et al., 2018; Ha & Schmidhuber, 2018; Hafner et al., 2020; Schrittwieser et al., 2020).

Another important factor is compositional reuse of learned modules, like prediction models and action policies. The cerebral cortex is composed of multiple areas for distinct representations of sensory, motor, and cognitive information. They form multiple cortico-cerebellar and cortico-basal ganglia loops for prediction and valuation using multiple representations (Doya, 1999; Samejima & Doya, 2007). Models based on the laminar architecture of the cortex with learnable connections between layers have reproduced behavioral data in diverse areas including visual pattern processing and working memory (Grossberg, 2021). While architectures for learning and combining multiple modules for prediction and control have been proposed (Dayan & Hinton, 1993; Dietterich, 2000; Doya et al., 2002; Morimoto & Doya, 2001; Sutton et al., 1998; Wiering & Schmidhuber, 1998; Wolpert & Kawato, 1998), how such architectures can be adaptively constructed is still an open issue. In the brain, there appear to be mechanisms for flexible "pathway gating", for example, by gating of inputs and outputs by inhibitory neurons (Wang & Yang, 2018) and by synchrony and temporal coherence (Palmigiano et al., 2017). In AI applications, selection and combination of pre-learned modules are mostly accomplished by human designers, but automating such processes is an important new direction (Hutter et al., 2019).

Last, but not least, learning to learn, or meta-learning, is also an important topic in neuroscience and AI research (Thrun & Pratt, 1998). One aspect of meta-learning is to adjust hyper-parameters of learning algorithms, such as the learning rate, temperature for exploration, and temporal discounting factor, to match characteristics of the environment and the stage of learning. Neuromodulators like acetylcholine, noradrenaline, and serotonin have been suggested to regulate these parameters of learning (Doya, 2002, 2008). Brain-inspired algorithms have been proposed to regulate those parameters based on the uncertainty of sensory observations and environmental dynamics (Parr & Friston, 2017; Yu & Dayan, 2005) and the amount of time available for decision, action, and learning (Doya et al., 2021; Kurth-Nelson & Redish, 2009; Reinke, 2018). Another aspect of meta-learning is finding relevant features for solving similar tasks (Courville et al., 2006; Gershman, 2015). By training a recurrent neural network for multiple tasks, hidden neurons can capture dynamics and parameters of a series of tasks for efficient re-learning, sometimes even without synaptic learning (Wang et al., 2018; Yang et al., 2019).

Further understanding of mechanisms of data-efficient learning in the brain would accelerate development of more adaptive

AI, even though there may not be off-the-shelf recipe to apply. Development of more flexible AI algorithms and architectures may also provide computational models to understand the brain mechanisms.

3. Social benefits of AI and neurotechnologies

Novel technologies developed through interactions of AI and brain science, such as deep learning, are already impacting society in areas such as information services, manufacturing, and healthcare, and they will deeply affect the ways people work, live, and think.

Although the potential benefits of AI and neurotechnologies are endless, here we review their ongoing contributions to sustainability, healthcare, and basic science.

3.1. AI for sustainability

Preserving and restoring the earth environment in the face of climate change and environmental pollution is a major challenge for society today. AI and deep learning can help tackle climate change in multiple ways (Rolnick et al., 2019). One obvious way is through more accurate and reliable prediction of the effects of human activities on the environment by integrating big data from satellites, weather stations, simulation models, and so forth (https://www.climatechange.ai/events/neurips2020). Another way is by helping to optimize device and material designs for renewable energy sources, such as solar power generation, energy harvesting, and efficient energy storage with batteries and other media. A complementary direction is optimization of energy usage. For example, machine learning has been applied to improve the efficiency of data center cooling by 40% (https://deepmind.com/blog/article/deepmind-ai-reduces-google-data-centre-cooling-bill-40). Partly due to such optimization, world-wide energy use by data centers has plateaued in the last decade despite ever-increasing demands (Masanet et al., 2020).

3.2. AI for healthcare

Medical diagnosis has been a classic target of AI research, such as the MYCIN project in the 1970s (Buchanan & Shortliffe, 1984). While classic expert systems relied on human-curated rules, modern expert systems like WATSON exploit data mining and machine learning to extract knowledge from vast collections of documents that no single human being could read in a lifetime (Xu et al., 2019). Furthermore, machine learning from big biomedical datasets, such as genomics and brain imaging data, has been utilized for diagnosis and prognosis of various diseases (Esteva et al., 2017; Uddin et al., 2019).

In the face of the COVID-19 pandemic, AI researchers also tried to apply machine learning to diagnose the disease from X-ray or CT images before PCR testing became widely available (Desai et al., 2020; Jamshidi et al., 2020). Prediction of the spread of infection and numbers of patients and fatalities under alternative sets of prevention measures is also another important domain in which models based on big data are being utilized (Arik et al., 2020).

Development of a new drug requires screening of numerous candidate structures, which can take many years. Docking simulations of potential pharmaceuticals with target molecules can accelerate such searches, but requires knowledge of the 3D structures of target molecules. Predicting protein structures from gene sequences has been regarded as a grand challenge in biology. A recent approach based on deep learning, AlphaFold (Senior et al., 2020) was recognized as the first practical solution to this challenge at the CASP-14 competition (Jumper et al., 2021).

3.3. The Nobel Turing Challenge

How can AI promote scientific discoveries? The Nobel-Turing Challenge (Kitano, 2016, 2021) is for an AI system to make a scientific discovery worthy of a Nobel Prize, by 2050, without the Nobel Committee's realizing that it was from an AI system. This is the Turing test at the Nobel Prize-level scientific activities by AI systems.

The challenge has two sub-goals. One is to determine what it takes to achieve a major scientific discovery and how we can understand and reproduce the process of scientific discovery. AI systems may find an alternative approach to scientific discovery, very different from what humans are doing now.

The second goal is to see how an AI system can become like a human scientist, autonomous enough to be able to choose the research topic and to be able to explain and communicate the findings to other members of the community. Such a system must be able to justify why a specific topic needs to be pursued and to propose extending collaborations. The second goal essentially implies that the machine must be able to align its value system with ours.

Although high-throughput machines and sophisticated analyses are available, scientific discoveries remain at a pre-industrial revolution stage. After researchers gather all the data, they scratch their heads and ponder the meaning. This process has changed little and the major challenge is how to automate the process of scientific discovery. Scientific discoveries are most often based on scientific intuition, or by serendipity of chaining unrelated events by analogy (Dunbar, 2000), or by accident, but how can these be defined computationally? How can this process be implemented in machines? That is a very important challenge now.

More than 20 years after it was established (Kitano, 2002), systems biology is facing the challenge of massive data and complexity. Systems biology is the science of AI, or AI–human hybrid systems. Mathematical frameworks have been developed to extract mathematical models underlying observed data (Bongard & Lipson, 2007; Brunton et al., 2016; Schaeffer, 2017).

There is, however, an information horizon problem. More than 2 million papers are published every year, or more than 4,100 papers per day, which is beyond the limit of human cognition. For example, we created the yeast signaling pathway 10 years ago by reading 1,500 papers (Kaizu et al., 2010), but it is almost impossible to update it by unaided human effort. There is also the issue of minority reports: How do we evaluate a small number of studies that conflict with the majority? Can such results simply be dismissed, or could they represent new discoveries for specific conditions? Another problem is human cognitive biases, such as anchoring bias, confirmation bias, and premature closure (Kahneman, 2011).

The fundamental premise of the Nobel-Turing Challenge is to create an engine for scientific discovery. Machine discovery is not a new concept. One of the first systems was Dendral, created in the late 1960s by Feigenbaum for discoveries in organic chemistry (Lindsay et al., 1980). Most recently, AlphaFold achieved atomic-level prediction of protein 3D structures from amino acid sequences by deep learning of known protein structures in a public database (Jumper et al., 2021; Jumper & Hassabis, 2022). Moreover, the "robot scientist" by Ross King creates hypotheses in yeast genetics, generates experimental protocols, runs fully automated robotic experiments, and performs data analysis to close the loop for new hypothesis generation (Coutant et al., 2019; King et al., 2009).

3.4. Processes of scientific discoveries

What are the major challenges to further progress in scientific discoveries? AI can redefine scientific discovery by

performing massive searches and verification in a hypothesis space, but it cannot search in an exhaustive manner, so how can the search be structured? Can there be a computational definition of serendipity?

Let us consider two examples. The first is the discovery of iPS cells by Shinya Yamanaka, who was awarded a Nobel Prize in Physiology and Medicine in 2012 (Takahashi & Yamanaka, 2006). From the FANTOM database, Yamanaka and colleagues discovered 24 candidate genes that can initialize a cell. They then performed single-gene knock-out experiments to identify the four Yamanaka factors necessary to induce pluripotent stem cells. This was accomplished by search and optimization.

The second example is the discovery of a conducting polymer thin film by MacDiarmind and Heeger, who were awarded a Nobel Prize in Chemistry in 2000 (Shirakawa et al., 1977). This discovery happened with the fortuitous formation of a polyacetylene thin film when a student accidentally used a reagent at 100 times higher concentration than intended. The researchers then optimized conditions for creation of polyacetylene thin film and for electrical conductivity. This illustrates a case of accident, search, and optimization.

The question is, "How can a machine reproduce processes like these, or achieve them by an alternative process?" There can be several strategies for this Nobel Turing Challenge. One is for capable research labs to harness AI assistants to facilitate scientific discovery. The GARUDA platform connects databases, simulators, and other tools, and is being used to predict heart arrhythmias and to perform multi-omics data analysis (Ghosh et al., 2011). In the ERATO-AMED project on influenza drug discovery, a pipeline has been established to extract knowledge from the literature, construct network models, and control theoretical analyses of critical nodes in drug design, such as searching for lead structures, performing docking simulations, and integrating humans and robots in the lab for experimental verification.

3.5. Science by AI and by humans

For an autonomous AI scientist, having a single, fixed pipeline is not enough (Kitano, 2016). The success of AlphaGo was based on hypothesis generation using a policy network and verification by a Monte-Carlo tree search using a value network (Silver et al., 2016). AlphaGo started with large-scale data from human competitions and explored billions of simulated self-plays until it was ultimately able to defeat human competitors. AlphaGo Zero did not use human data, but by further refinement, it became more adept than either humans or AlphaGo (Silver et al., 2017). What we learned from this is that humans play games using strategies that humans can understand and with which they are comfortable, but machines explore the game space in ways that humans never imagined and at which they are much more skilled (Fig. 1 A).

In the game of Go, simulation is very precise and efficient, but achieving high-precision experiments in biology in combination with AI/ML and modeling/simulation is a much bigger challenge. One significant conundrum is how to handle negative data, which are important for machine learning. Existing papers and data tend to be biased toward positive results, which may be even harmful for machine learning algorithms.

Many people say that asking the right questions is essential for scientific discovery and doubt whether AI can ask the right questions (Choe & Mann, 2012). But this may be due to human cognitive limitations and sociological limitations scientists face to make discoveries during their research careers. However, if this bottleneck can be eliminated with greatly accelerated hypothesis-testing cycles so that every question can be answered quickly, asking the right questions may no longer be so important. Asking only the "right" questions from a human perspective may actually be suboptimal (Fig. 1B).

Massive searches and verification of hypothesis space may redefine scientific discovery. One issue is whether there should be one universal AI scientist, or a population of specialized AI scientists collaborating with each other. The latter is similar to what human scientists do and may produce results in the near future, together with human scientists, but specialized AI scientists may make big discoveries that human scientists cannot make.

At the end of the day, we may have a better understanding of what serendipity means and what constitutes scientific intuition. With AI scientists, we would probably find alternative forms of scientific discovery at unprecedented speed. In creating autonomous intelligence that can evolve by itself, we also need to address ethical concerns.

4. Risks of AI and neurotechnologies

As we have seen above, the joint progress in AI and neuroscience has transformed super-human AI agents from science fiction to reality. However, this progress raises social concerns in ethics and governance. Norbert Wiener wrote (Wiener, 1960):

> *We had better be quite sure that the purpose put into the machine is the purpose which we really desire.*

When we create something with a purpose, we must also consider the risk of unintended use or failure to achieve the purpose. Because we are deeply engaged in science and technology, we are distressed when science and technology are misused.

Progress in science and technology has opened new horizons to human civilization, and in return, it has caused various incidents and accidents. The Golem book series by Collins and Pinch (Collins & Pinch, 2012, 2014) discussed various incidents in which technologies were misused, or spiraled out of control, including the Chernobyl disaster and "Climategate". They also question the structure of society and the responsibility of the experts who created the Golems. This is why we should consider not only the risks of AI and neurotechnologies, but also responsibilities of researchers and engineers in AI and brain sciences.

There are different types of risks that we need to consider. One is intentional abuse by humans, attacking other humans, due to greed, ambition, or hatred. Another type is unexpected malfunctions or side effects of technologies, despite their intended use for good. Beyond these problems are indirect effects on society, e.g., a technology benefiting some, but disadvantaging others. Below we examine different types of risks associated with AI and neurotechnologies.

4.1. Intentional abuse

In addition to unintended disasters resulting from new technologies, we must worry about intentional abuses. "Deep fake" pictures and movies are an example of how AI can empower people with mischievous or malicious intent. Another case is Dr. Evil in Austin Powers movies (Roach, 1997), who is not interested in benefiting human beings, but employs AI systems that will enable him to take over the world. Misuses of AI are an extension of the problem we already have with cybercrime, which is already a trillion-dollar problem, but will become much worse because we may lose control over how AI systems perform. And then the consequences for humanity will be very serious.

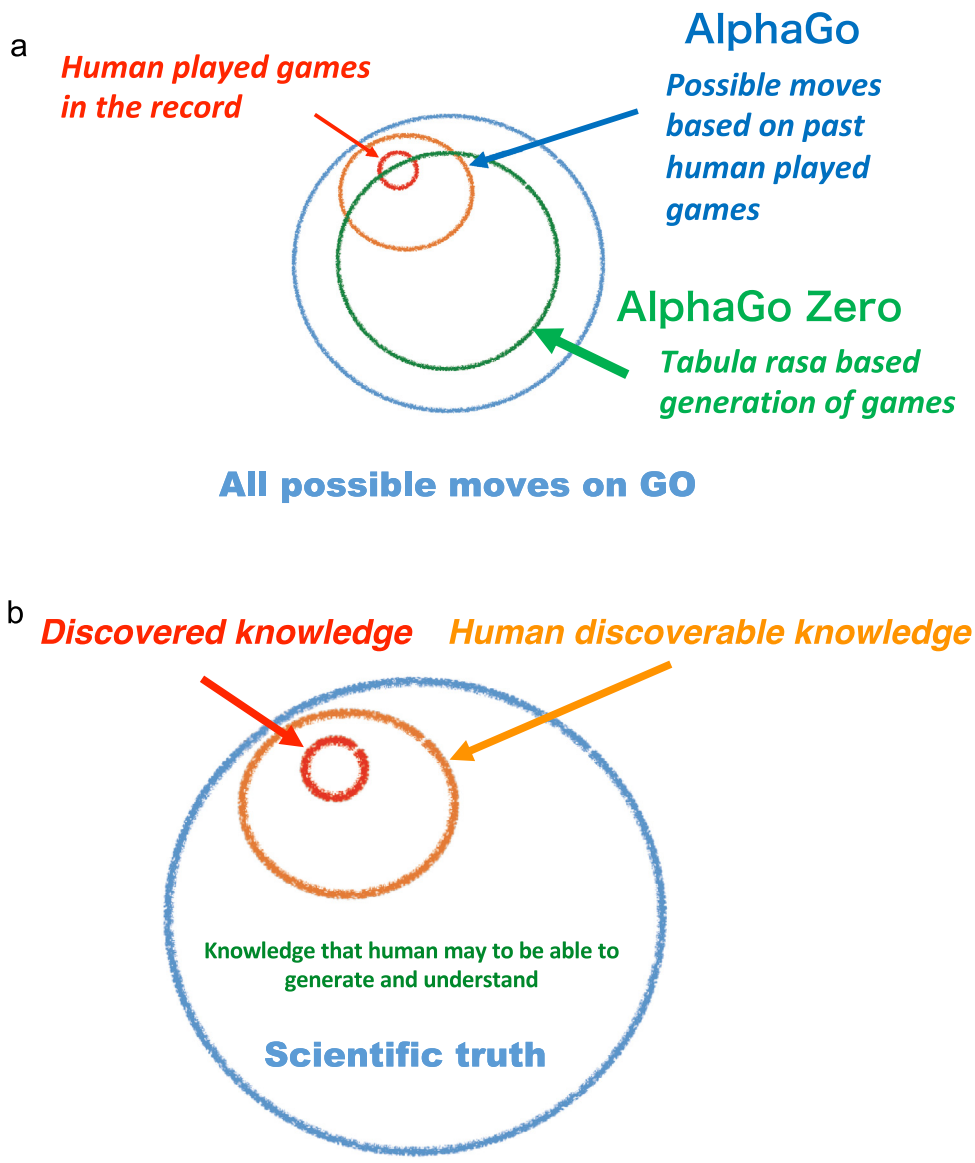

Fig. 1. (A) AlphaGo (Silver et al., 2016) started with large scale data of human Go playing and then explored billions of simulated self-play to ultimately defeat humans. AlphaGo Zero (Silver et al., 2017) did not use human data and with further refinement, became more capable than human competitors and Alpha Go. Machines can explore the game space in ways that humans never imagined and can actually become more capable. (B) Questions humans can ask are limited by their cognitive capacity and research careers. If super-fast hypothesis-testing can answer every question quickly, asking the right questions may not be so important for AI scientists.

4.2. AI in battlefields

Among many risks of AI, its use in autonomous lethal weapons is the most imminent and serious (Russell, 2015). We can build robots that can replace humans. And this raises dilemmas when it comes to robots in the battlefield. There is a risk of lower restraint in attacking humans if machines are to make decisions. Currently there are robots based on self-flying drones and anomaly detection, but none of them are as flexible or versatile as humans yet.

There is an account of how soldiers fought the Taliban (Scharre, 2018). A girl of five or six years headed out of her village herding goats. She left the goats and then Taliban fighters arrived soon after. The laws of war do not specify a lower age limit for combatants, who are classified based on behavior. If the girl was spotting the enemy for the Taliban, then she was a lawful target for elimination. Of course, killing her would have been wrong, morally, if not legally, but what would a machine have done in this case, if it had been programmed to kill enemy combatants? It would have attacked the girl. Humans have a responsibility to carefully consider what decisions robots should be permitted to make.

Machine–human relationships are also important. There is a risk of humans anthropomorphizing AI robots, and misplacing rights and responsibilities. In a military experiment in which a mine sweeping robot had its limbs blown off, a soldier requested that the experiment be halted on grounds that the test was inhumane (Washington Post, 2007). This example suggests that people are likely to misjudge robotic capacities in other ways, and clearly illustrates the importance of considering moral agency and responsibility in the use of AI in war.

This problem is not only the responsibility of researchers directly involved in military research, because civilian technologies can often be diverted for military use. Technologies for

autonomous vehicles or medical care can also be used in military affairs. This is why researchers not directly in military projects need to remain vigilant in regard to ethical questions.

4.3. Brain–computer interfaces and cognitive enhancement

With advances of electronic devices and machine learning, it is becoming feasible to extract brain signals from electrodes over the skull or implanted in the brain. While the brain computer interface (BCI; also called brain machine interface, BMI) can yield huge benefits for patients with spinal cord injuries or amyotrophic lateral sclerosis (ALS), it also raises security and ethical issues. As all online computers globally are subject to cyber attacks, how can we avoid brain hijacking by attacks to brain-connected computers? And even if safety and reliability of BCI can be established, there will be attempts by healthy people to use it for cognitive enhancement, which can create ethical and legal issues regarding agency and responsibility.

A discussion by a group of neuroscientists and AI researchers, Morningside Group, identified four ethical priorities in BCI: privacy, identity, agency, and equality (Yuste et al., 2017). Subsequent discussions at the Global Neuroethics Summit that brought together researchers from the International Brain Initiative (IBI, https://www.internationalbraininitiative.org) identified five major questions for neuroscientists (Global Neuroethics Summit Delegates et al., 2018):

Q1. What is the potential impact of a model or neuroscientific account of disease on individuals, communities, and society?

Q2. What are the ethical standards for biological material and data collection? How do local standards compare to those of global collaborators?

Q3. What is the moral significance of neural systems that are under development in neuroscience research laboratories?

Q4. How could brain interventions impact or reduce autonomy?

Q5. In which contexts might a neuroscientific technology/innovation be used or deployed?

Their perspective paper considers practical issues in the context of multiple cultural backgrounds, such as the individual versus society and the brain versus the body (Wang et al., 2019).

4.4. AI affects people's decisions

Even without electrodes in the brain, AI is already affecting how people perceive the world and make decisions. AI classification and recommendation programs trained by human outputs can learn existing biases and inequity and can reinforce them or make them persistent. Efforts have been made to remove unethical biases from training data sets or learned results, but there are even deeper issues in interactions between AI and users.

Content selection algorithms in social media choose news articles and videos for people to read and watch, which occupy much of the time of billions of people every day. These algorithms are designed to maximize click-through, or some other proxy, such as engagement and attention. In order to maximize click-through, they send us material of interest to us, which admittedly, is better than sending us material of no interest.

But in fact, that is not just what these algorithms do (Benkler et al., 2018; Reich et al., 2021). They do not simply learn what people want. Reinforcement learning algorithms execute actions in order to change the state of the world so as to maximize their reward. In this case, the state of the world is a human brain. In essence, what the algorithms do is to modify people to be more predictable in their clicking behavior. They do that because greater predictability makes it easier for the algorithm to send people content that they will select. In so doing, the algorithm is simply maximizing its objective, but it has no notion that the human users might become an eco-terrorist, a neofascist, or any other kind of extreme, but predictable person (Carroll et al., 2021; Russell, 2020). There are already some studies on how to avoid such problems (Stray, 2020; Stray et al., 2021).

4.5. Social gaps and human enfeeblement

As Andy Clark notes, people are already natural-born cyborgs (Clark, 2003). In addition to dependence upon glasses to supplement vision, people rely on digital spaces for their memories and information processing. Scheduling and communication records are stored in digital spaces. Research, writing, and meetings can be no longer be accomplished without information services and devices. Living in a modern society means living as a cyborg.

A popular fear about AI and robotics is that they steal jobs from humans, leaving people unemployed. The worst scenario is that thoughtful human experts are replaced by simple AI agents that can handle routine tasks quickly. However, an opposing scenario is that AI and robotics empower people in many jobs and create new businesses, so that people can enjoy life, less burdened by simple labors. What is actually occurring is a mix of both, causing disparities in society with people who create AI, those who utilize AI, and those who can do neither and just wind up as data sources for AI. Such disparities can cause huge inequalities within and between countries, which may exacerbate social and international tensions.

As AI becomes more advanced and autonomous, the need for humans to create AI may diminish. This leads to what is called the "Wall-E problem", after a movie in which AI systems were designed to run civilization on behalf of humans. In such a scenario, humans would no longer have an incentive to understand how their own civilization works. They would simply be beneficiaries or recipients. Essentially they would be passengers on a cruise ship on a cruise that continues forever. This is a disastrous future. It is a cultural problem because we may concede management of civilization not to the next generation of humans, but to the next generation of machines. We have to understand how societies can maintain cultural and intellectual vigor in spite of increasingly prevalent AI. At the moment, these problems are completely unresolved, and new ideas are needed to solve them. To protect the dignity and value of human beings while advancing technology and science, multi-stakeholder discussions in various fields such as philosophy, law and society are required.

5. Ethics and governance of AI and neurotechnologies

Research in artificial intelligence and neuroscience that enables human-like or super-human intelligence also raises questions regarding the interface between science, technology, and society. It is a responsibility of scientists to imagine how their research will be applied to society. Discussing ethics and governance upstream of the research is really important. Responsible research innovation (RRI) is a key term for asking what kind of society we wish to inhabit. This is not a question that can be answered easily. Therefore, we must continue dialogue with the general public.

In fact, many AI researchers have already taken those risks seriously and initiated such discussions. For example, the ethics committee of the Japanese Society for Artificial Intelligence (JSAI) issued the following guidelines in 2017 (http://ai-elsi.org/archives/514):

1. Contribution to humans
2. Compliance with laws and regulations
3. Respect for the privacy of others
4. Fairness

5. Security
6. Action with integrity
7. Accountability
8. Communication with society and self-development
9. Compliance with ethical guidelines by AI itself

The unique feature of these guidelines is Article 9, implying that AI itself must abide by the policies described above in the same manner as members of the JSAI, in order to become quasi-members of society. There is a question whether AI will be given basic human rights, such as freedom of expression and unsupervised autonomy, if it becomes a human partner. Members of the JSAI ethics committee added this article to promote a deeper discussions on the ideal form of artificial intelligence.

The Asilomar Conference in 2017 concluded with 23 principles for safe governance of AI technologies (https://futureoflife.org/ai-principles/). In 2019, the EU's expert group on AI presented Ethics Guidelines for Trustworthy Artificial Intelligence (https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai). Academic, industrial, and social organizations formed a non-profit organization Partnership on AI (https://partnershiponai.org) which aims to create actionable programs for the AI community.

The meta-analysis of AI guidelines by many governmental and non-governmental organizations identified convergence to five ethical principles: transparency, justice and fairness, non-maleficence, responsibility, and privacy (Jobin et al., 2019). Now the important question is how these principles can actually be implemented in a reliable and sustainable way (Hickok, 2020; Mittelstadt, 2019).

5.1. Human compatible AI

In 1950, Alan Turing, the founder of computer science and arguably also the founder of artificial intelligence, published a very famous paper, "Computing Machinery and Intelligence", which laid out some of the basic ideas of AI and defined what we now call the Turing test (Turing, 1950). He also predicted the application of machine learning as the best way to build AI systems, among many other very innovative ideas.

What people are less familiar with a talk that he gave on the radio in 1951, in which he said,

> *It seems probable that once the machine thinking method had started it would not take long to outstrip our people powers. At some stage therefore we should have to expect the machines to take control.*

This was said with almost a sense of resignation. He offered no solution. It was, and continues to be, a very pessimistic prediction about the future.

What we need to do today is to explain why this problem arises and why it is the case that better AI leads to worse outcomes. We have already considered scenarios in which greedy or hateful people use AI technologies to exploit or attack other people. However, even among people with no intent to harm others, AI can still cause problems in trying to serve optimally. A new book, entitled, "Human Compatible Artificial Intelligence" (Russell, 2020) analyzes how such problems occur and proposes how we can avoid them in order to develop AI that is compatible with human existence.

We already have self-driving cars and AI systems have beaten human champions in games like Go and chess, which were previously regarded as the pinnacle of human intellectual achievement. As this progress continues, AI systems will eventually make better real-world decisions than humans. The capacity to make sound, wise decisions gives us power over the world. Knowledge and intelligence are power.

What we are going to do is to build systems that are more powerful than ourselves. So the question that Turing asked is, "How do we retain power over entities more powerful than ourselves, indefinitely?" Turing had no answer for that question, and that was the reason for his pessimism. Stephen Hawking made a similar point, in an editorial that was co-authored with Russell, Tegmark, and Wilczeck in 2014 (Hawking et al., 2014):

> *Success in creating AI would be the biggest event in human history. Unfortunately it might also be the last, unless we learn how to avoid the risks.*

That editorial observed that it is possible to escape the path that Turing predicted. To do that, we have to understand the source of the problem. The source lies in the way we think about artificial intelligence, which we call the *standard model* of AI. The standard model is to create machinery that achieves or optimizes objectives that we specify. The objective is defined by humans and plugged into the machine, which then finds a way of solving it. In fact, one could say it is the standard model underlying much of the technical innovation of the 20th century.

For example, if one tells a self-driving car "take me to the office by 9 am" or "as soon as possible", the self-driving car figures out how to do that. The car simply adopts the destination and time as its objective. This is how we build all AI systems today. The problem is that when we start moving out of the lab and into the real world, we find that we are unable to specify objectives completely and correctly. For example, what if there is a traffic jam or snowfall?

This is not a new point. In fact, we've known this for thousands of years. For example, in the ancient Greek legend, King Midas asked the gods that everything he touched might turn to gold. This was the objective he specified, and the gods granted his objective. They were the optimizing machinery. Of course, as the story goes, his food, his drink, and his family all turned to gold, and then he died in misery of starvation. Then there are all those tales of the Genie in the Lamp. When the possessor of the lamp rubs the lamp, the Genie grants him three wishes. What is the third wish? It's always, "Please undo the first two wishes because I've ruined the world!" In fact, many cultures have the same basic legend or story, just to remind people that getting what one asked for is not always what one really wants.

Unfortunately, with the standard model of AI, we get what we ask for. A modern example is what has happened with social media, as presented above. The algorithm just wants the user to be predictable, and that's what it does to the user. What we see from this lesson is that with the standard model and poorly specified objectives, the better the AI system, the worse the outcome. That is, the more capable the AI system, the more it will be able to mess with the world in order to achieve this incorrectly specified objective. In a sense we are setting up a chess match between ourselves and the machines with the fate of the world as the prize. We don't want to be in that chess match.

A possible alternative to maximize the objective is to achieve it passably well. However, setting a particular threshold can also be problematic under uncertainty, e.g., driving to the office by a given time despite traffic jams or snowfalls. Another proposal is "quantilizers" (Taylor, 2016), which choose an action randomly from those that achieve above a given percentile threshold. The quantilization approach is robust under some restrictive assumptions, but can have arbitrarily bad downside risk in general.

5.2. New principles for provably beneficial AI

These arguments suggest the need for a new model for artificial intelligence, or what Russell has called "Provably Beneficial

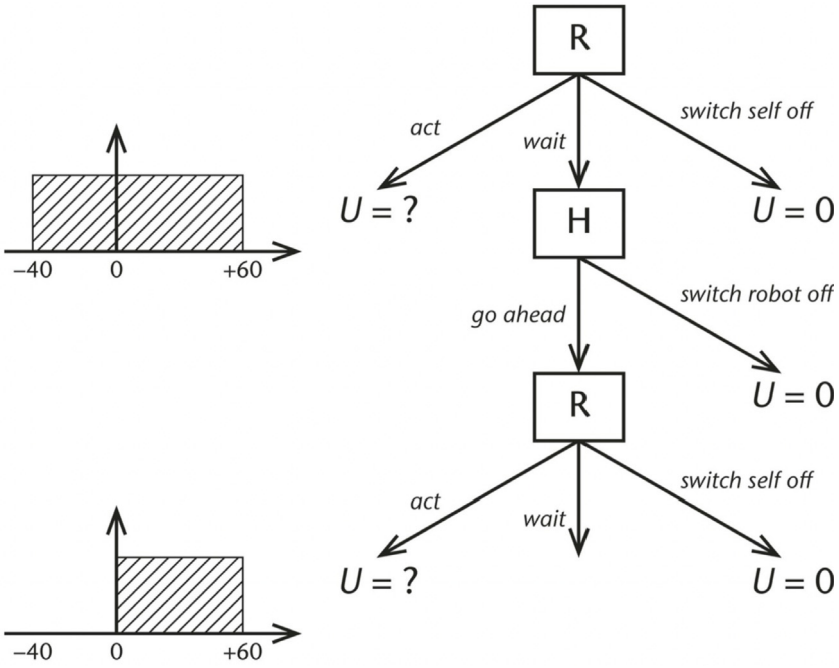

Fig. 2. In the "assistance game" setting (Russell, 2020), the goal of the robot (R) is to satisfy uncertain preferences of the human (H) by learning them through interactions with humans. When the utility (U) of an action may be negative, the robot is motivated to ask permission, or switch itself off, rather than just acting.

AI" (Russell, 2020). If we design AI according to the right principles, they will necessarily be beneficial to us and we will avoid the problem of misspecified objectives.

There are three principles. The first is that a robot's only goal is to satisfy human preferences. Here, human preferences do not just mean what kind of pizza we like. It is our preferences about everything we care about for the entire future. And it is not just one individual's preferences, but the preferences of all human beings, which may vary enormously among people in different societies.

The second principle is that a robot does not know what those preferences are. And the robot knows that it does not know what those preferences are. This uncertainty enables us to design AI systems that are safe, as explained below.

The third principle is that the robot learns about human preferences from the evidence provided by our choices. We do not provide complete evidence about our preferences. It is imperfect evidence because we don't behave perfectly rationally. Our behavior doesn't perfectly reflect our true underlying preferences. Nonetheless, it provides evidence about what those preferences are.

We can turn those three principles into what we call an "assistance game" (Shah et al., 2020, 2019). This is a formal mathematical framework that defines the problem the AI system is to solve. It is a game because it has multiple participants, humans and machines, and incentive systems. The entire purpose of the machines in this game is to assist humans. And when we look at the solutions of these assistance games, we find that the robot necessarily defers to the humans and follows their instructions, because those instructions convey information about preferences. When the robot is not sure whether its plan is consistent with our preferences, it is motivated to ask permission before carrying out the plan (Fig. 2). These desirable properties all arise from having a sufficiently broad prior belief about human preferences. If the robot's belief assigns probability zero to all preferences that are acceptably close to actual human preferences, then it is essentially operating according to the standard model and its behavior may be problematic. On the other hand, it makes sense to build in prior probabilities that assign *higher* probabilities to preferences that plausibly correspond to what most or nearly all humans prefer. For example, most humans probably prefer better health to worse health, all other things being equal.

5.3. The off-switch problem

In an extreme case, a robot that is solving an assistance game will allow itself to be switched off. And this is really the core of the control problem. If you can't switch the robot off, then it's "game over". With this new approach to AI, the better the AI, the better the outcome, because the AI system is better able to learn our preferences and it is better able to assist us in achieving them.

If we do things the classical way according to the standard model of AI, we give the robot an explicit goal, for example, to fetch the coffee. So now the robot says to itself "I must fetch the coffee. I can't fetch the coffee if I'm dead. Therefore, I must disable my off-switch". And perhaps it will drive away all the customers at Starbucks, just to be certain it can return with the coffee in time. This is how a robot that is programmed according to today's standard model of AI might behave, and this is what we want to avoid in the new model.

When there is uncertainty about the objective, even though the robot may know that you want coffee, it may be uncertain about all your other preferences (for example, your preferences regarding the well-being of other customers at Starbucks). This means that the robot will think to itself in a very different way: "A human might switch me off, but only if I'm doing something wrong".

According to the first and second principles, the robot knows that it doesn't know what is wrong, but it knows that it does not want to do anything wrong. This actually gives the robot an incentive to allow itself to be switched it off. In fact, it *wants* to be switched off, rather than to do something wrong.

We can formulate this informal argument mathematically, and prove the theorem that a robot that is designed this way is

demonstrably beneficial to humans and that it will allow a human to switch it off. It has a positive incentive to do so, as long as it is uncertain about the human's preferences. There is actually much research needed on this new framework, some of which has already been done (Fickinger et al., 2020; Hadfield-Menell et al., 2017; Shah et al., 2020, 2019), but more still needs to be done to put this new model into practice (Christian, 2020; Gabriel, 2020).

One direction we must pursue is to change AI from operating on behalf of a single individual to operating on behalf of mankind. Making decisions on behalf of many people has been studied for thousands of years in moral philosophy, from Aristotle and Plato to Bentham and Mill for the theory of utilitarianism and modern economics. Tradeoffs are inevitable – after all, not everyone can be Ruler of the Universe – and utilitarian theory describes a normative criterion for such tradeoffs. Other moral theories posit further constraints that may override or replace utilitarian principles. Foe example, "sadistic" preferences, whereby one person derives utility solely from the suffering of another, should be excluded.

We also have to confront the fact that humans are not perfectly rational and our behavior doesn't always perfectly reflect our underlying preferences. In order to understand how humans behave in providing evidence of their underlying preferences, we have to look to cognitive psychology and neuroscience to build models of human cognition and then to invert the models to ascertain the underlying preferences from observed behaviors.

On the practical side of AI, we suggest that the AI community will have to rethink theoretical foundations. The present theory of AI is founded on the standard model, in which it is assumed that the objective is completely and perfectly known. We now see that this assumption is false. Each area of AI will need to be rebuilt, including search algorithms, planning algorithms, reinforcement learning, and natural language understanding. We can also examine applications that really demonstrate the value of the new model. For example, self-driving cars need to understand the preferences, not only of the passenger, but also of pedestrians and of all other cars on the road. The AI system controlling a vehicle should make decisions given its understanding or lack thereof, of the preferences of other vehicles (their drivers and passengers) as well as the preferences of its own passengers. These are the kinds of AI systems that we would really like to build and that we would like to own, because they will do what we want. This is both possible and desirable.

5.4. Autonomy and sociality

Most machine learning algorithms are formulated to minimize or maximize an objective function. A reinforcement learning agent learns to maximize the reward acquired, but designing an appropriate reward function is by no means trivial.

In a series of works using "Cyber Rodents", it has been demonstrated that reward functions to promote survival and software-reproduction can be acquired by embodied evolution within a colony of robots (Elfwing et al., 2011) (Fig. 3). In some of these robot colonies, robots adopted different reproductive tactics (Elfwing & Doya, 2014). There have also been studies on "intrinsic motivation", allowing agents to have curiosity to promote directed exploration (Oudeyer, 2018).

These studies have already demonstrated that AI agents can set their own goals as reward functions to survive (to avoid being discharged or switched off), to reproduce (to proliferate via copies), and to acquire information (which can contribute to the former). This may sound like a dream, but some people find it dangerous.

AI agents that can set their own goals and learn to accomplish them may be fantastic. They may make scientific discoveries, deliver new technologies, start new industries, and create new cultures, but at the same time, we must be very careful about the risks they pose, as presented in previous sections.

Even if such creative AI agents perform perfectly, they may have side effects that are hard to predict. Some of them may behave irregularly or become uncontrollable with some parameter setting or in unexpected situations, just as humans suffer neuropsychiatric disorders. However, the most immediate danger is that autonomous AI agents can be exploited by extremists to satisfy greed, ambition, or hatred.

In addressing those risks, learning from human society and from the brains that create the society may be very important. This is because, even though autonomous AI agents can be dangerous, humans are also very dangerous, arguably the most dangerous species on this planet. Mankind has already exterminated many species and may yet do the same to itself, by nuclear war or environmental destruction, thanks to its intelligence.

Nonetheless, human society has also created mechanisms to avoid such catastrophes. Democracy is one such mechanism to avoid over-concentration of power in any one person or group, which most often leads to resource misuse, tyranny, or war. In politics, we implement elections with term limits, separation of governmental powers, and local governance. In the economy, policies like anti-trust laws and the right to strike have been established to avoid over-concentration of economic power, which is not only unfair, but can also hinder innovation. In science, peer-review is intended to find any flaws in prominent theories or in new technologies, in order to promote sustained development.

In the near future, we would be working with varieties of AI agents with diverse expertise. In such a society, it would be desirable to apply the wisdom of democracy to AI agents. We should not allow just one AI program unlimited control over the world. We should have multiple, open-sourced AI programs with overlapping capabilities to provide services. Perhaps the most desirable way is to employ heterogenous AI agents that can explain their answers to form committees or for peer-reviewing mechanisms to avoid misuse of power (Montes & Goertzel, 2018).

In such a future society, how those AI agents might communicate with each other efficiently, but in a human-comprehensible way, is a fascinating research issue. Clonal AI agents might just transmit their hidden layer vectors in order to communicate, but that does not work among heterogenous, intelligent agents, including humans. AI agents should also have some kind of social skills to enhance their capabilities. A selfish agent should eventually be excluded from the working community.

There has been great progress in understanding the brain's mechanisms for social behaviors (Rand et al., 2012). Traditionally, evolutionarily older parts of the brain, like the amygdala, were supposed to govern animal-like, brutal behaviors, while newer parts of the brain, like the prefrontal cortex, are the source of rational behaviors. However, recent MRI studies on social orientation have revealed that pro-social people who favor equal division of rewards tend to show higher volume and activity in the amygdala, while sociopathic people, who just seek to maximize their own rewards have well-developed prefrontal cortices, activated during selfish decisions (Haruno & Frith, 2010; Yamagishi et al., 2016).

This suggests that before acquisition of high-performance prediction and planning capabilities, our brains had already evolved to care for others and to favor fair division of rewards (Levine, 2021). Perhaps in the design of operating systems or core libraries of AI agents, sensitivity to responses of humans and other agents should be central functions. This would help avoid runaway behavior like the coffee fetching robot mentioned in the previous section.

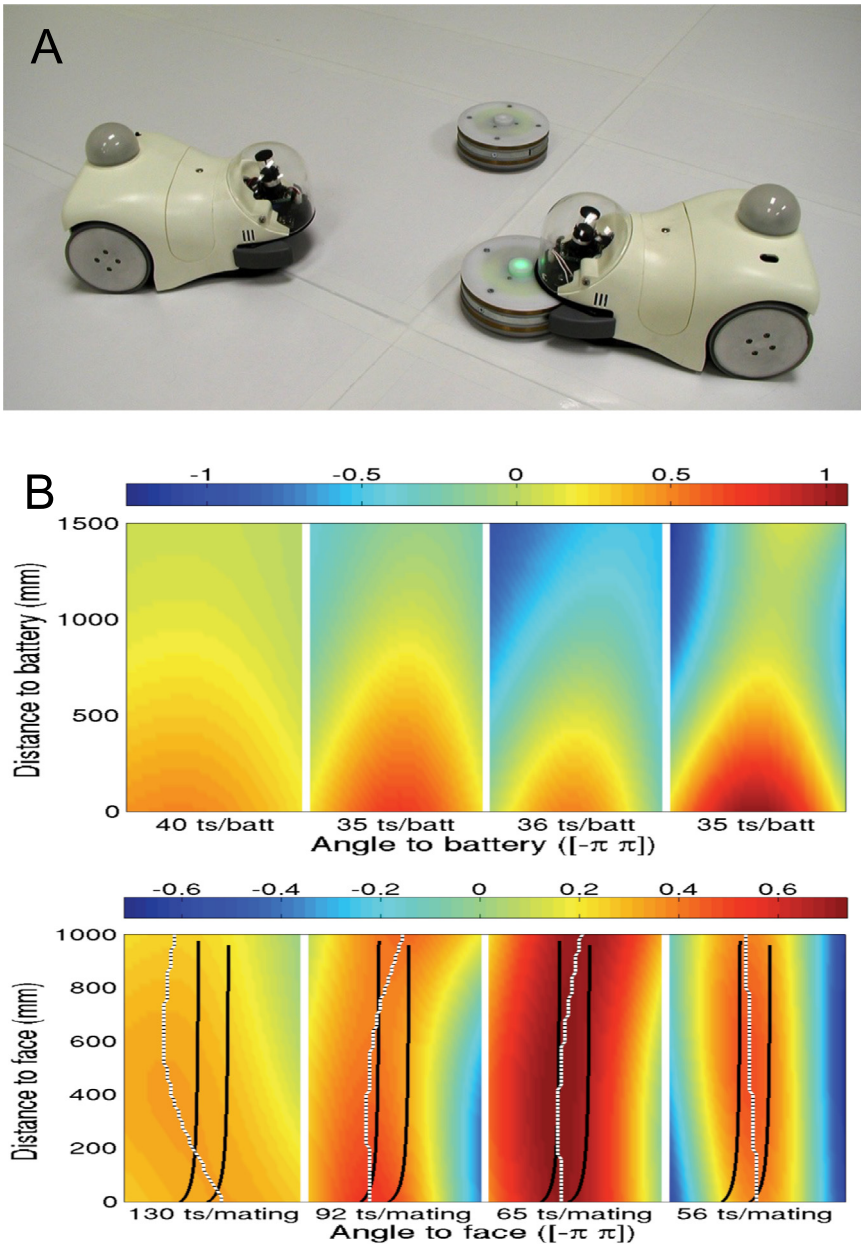

Fig. 3. (A) Cyber Rodent robots have two basic capabilities: recharging from batteries for survival and exchanging parameters of the program (genes) by infrared communication for software reproduction (Doya & Uchibe, 2005). When a robot depletes its internal battery without successful recharging, it is removed from the colony (death). When robots exchange genes by IR communication (mating), a small noise is added for mutation and adopted with a probability proportional to the sender's charge level. (B) Through this embodied evolution framework, distinct reward functions for the sight of a battery pack and for another robot evolved to promote frequent battery acquisition and mating (Elfwing et al., 2011).

6. Discussion

In this article, we first reviewed the co-evolution of AI and brain science and discussed three major ways that AI must emulate the brain: energy efficiency, data efficiency, and autonomy and sociality. We then examined benefits and risks of AI and neurotechnologies, including how AI can facilitate scientific discoveries, and types of problems new technologies can create, including intentional misuse, unintended consequences, and social disparities. We then considered ethics and governance of AI and neurotechnologies. One proposal is to let AI agents continually infer human preferences and another is to learn from democratic human society to prevent excessive consolidation of power. These proposals are rather orthogonal and may be integrated in the future designs of more autonomous AI agents.

6.1. Open issues in AI and neurotechnologies for society

Human Compatible AI argues that the standard model of AI, to optimize user-specified objectives will lead to a loss of human control as AI systems become increasingly intelligent and that we should adopt a new model of AI. If AI is designed to continually infer human preferences, it will prove beneficial to humans and will allow humans to retain control over it. One critical question regarding this paradigm is "What is human preference?" There are billions of humans with as many individual preferences. How can a machine make decisions on behalf of humanity? How should it trade off among the conflicting preferences of many people? One particularly difficult scenario involves persons who derive pleasure from the suffering of others. In cases of such

"negative altruism", AI must be designed to give no weight to destructive preferences.

Perhaps AI should develop a dynamic way to align itself with different value systems. Rather than uploading any specific value system into the AI system, the AI system should be constantly uncertain about what people want and how to trade off those preferences.

Human beings are imperfect, so mapping from true underlying preferences to actual behavior is very complicated. We need to be able to reverse engineer it to design AI systems. Also preferences are not stable and are obviously changeable, since we are not born with the complex preferences we develop as adults. AI systems may exploit their capacity to modify human preferences so as to make them easier to satisfy. An example is what happens in social media. Whether the change of human preference is good or bad is nevertheless an open question, since it depends on what the change is.

Another issue is how much autonomy or self-judgement we should allow AI agents. Can they explore online job markets for jobs that they can do well? Can there be a company run by AI agents to develop new businesses? Or should AI agents remain as servants or employees of humans? How much autonomy should AI agents be given? Perhaps there is no fixed solution and AI agents must learn continuously, just as humans do.

As in human organizations without diversity of opinion, bad decisions often result. An organization composed only of AI agents would not perform well. No matter how well such agents learn to infer human preferences, they will never do it perfectly. Except for simple, routine matters, human participation in decision-making should be mandatory. A practical question is how to guarantee such human auditing or supervision for an ever growing AI world.

6.2. Need for inclusive and diverse discussions

Many AI and neuroscience researchers are thinking about the basic questions of what is human. An emerging field called neurophilosophy addresses such issues by a diverse group of people. There is no "ethical prescription" for AI researchers, but it is clear that we should not continue business as usual in the AI community. We have to adopt a new technical foundation that will lead to better AI system designs. For people in industry, ignoring AI ethics and governance can risk the reputation and survival of the company. Education programs for university students to consider ethics and governance issues of artificial intelligence and robotics are being created and made accessible online. It is also important for researchers in the social sciences and humanities to have adequate understanding of AI and brain science in order to hold useful discussions about their risks. We are living in a society in which science and technology are transforming its infrastructure. This is why we must create a forum in which the general public can help to determine what technologies we utilize and how we manage their use.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by JSPS KAKENHI Grants 16K21738, 16H06561, 16H06563 and 16H06571, and research support of Okinawa Institute of Science and Technology Graduate University to KD. We thank Dr. Steven Aird for deep proofreading and editing of the manuscript.

References

Amari, S. (1967). A thoery of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers*, *EC-16*(3), 299–307.

Arik, S. O., Li, C.-L., Yoon, J., Sinha, R., Epshteyn, A., Le, L. T., Menon, V., Singh, S., Zhang, L., & Yoder, N. (2020). Interpretable sequence learning for COVID-19 forecasting. In *NeurIPS 2020*.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal Ganglia* (pp. 215–232).

Barto, A. G., Sutton, R. S., & Andersen, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, *13*(5), 834–846. http://dx.doi.org/10.1109/TSMC.1983.6313077.

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Bongard, J., & Lipson, H. (2007). Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(24), 9943–9948. http://dx.doi.org/10.1073/pnas.0609476104.

Boyan, J. A., & Moore, A. W. (1995). Generalization in reinforcement learning: Safely approximating the value function. In G. Tesauro, D. S. Touretzky, & T. K. Leen (Eds.), *Advances in neural information processing systems, Vol. 7* (pp. 369–376).

Bruce, C., Desimone, R., & Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology*, *46*(2), 369–384. http://dx.doi.org/10.1152/jn.1981.46.2.369.

Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(15), 3932–3937. http://dx.doi.org/10.1073/pnas.1517384113.

Buchanan, B. G., & Shortliffe, E. H. (1984). *Rule based expert systems: the MYCIN experiments of the stanford heuristic programming project*. Addison-Wisley.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, *37*(1), 54–115. http://dx.doi.org/10.1016/s0734-189x(87)80014-2.

Carroll, M., Hadfield-Menell, D., Russell, S., & Dragan, A. (2021). Estimating and penalizing preference shifts in recommender systems. In *4th FAccTRec workshop on responsible recommendation*. https://facctrec.github.io/facctrec2021/.

Choe, Y., & Mann, T. A. (2012). From problem solving to problem posing. *Brain-Mind Magazine*, *1*(1), 7–8. http://dx.doi.org/10.1145/130385.130417.

Christian, B. (2020). *The alignment problem: Machine learning and human values*. W. W. Norton.

Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford University Press.

Collins, H. M., & Pinch, T. (2012). *The Golem: What you should know about science* (2nd ed.). Cambridge University Press.

Collins, H. M., & Pinch, T. (2014). *The Golem at large: What you should know about technology*. Cambridge University Press.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294–300. http://dx.doi.org/10.1016/j.tics.2006.05.004.

Coutant, A., Roper, K., Trejo-Banos, D., Bouthinon, D., Carpenter, M., Grzebyta, J., Santini, G., Soldano, H., Elati, M., Ramon, J., Rouveirol, C., Soldatova, L. N., & King, R. D. (2019). Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(36), 18142–18147. http://dx.doi.org/10.1073/pnas.1900548116.

Dayan, P., & Hinton, G. E. (1993). Feudal reinforcement learning. In C. L. Giles, S. J. Hanson, & J. D. Cowan (Eds.), *Advances in neural information processing systems, Vol. 5* (pp. 271–278).

Desai, S. B., Pareek, A., & Lungren, M. P. (2020). Deep learning and its role in COVID-19 medical imaging. *Intelligence-Based Medicine*, *3*, Article 100013. http://dx.doi.org/10.1016/j.ibmed.2020.100013.

Dietterich, T. G. (2000). Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, *13*, 227–303, <Go to ISI>://WOS:000165572900001.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex. *Neural Networks*, *12*, 961–974. http://dx.doi.org/10.1016/S0893-6080(99)00046-5.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*, 495–506. http://dx.doi.org/10.1016/S0893-6080(02)00044-8.

Doya, K. (2008). Modulators of decision making. *Nature Neuroscience*, *11*(4), 410–416. http://dx.doi.org/10.1038/nn2077.

Doya, K., Miyazaki, K. W., & Miyazaki, K. (2021). Serotonergic modulation of cognitive computations. *Current Opinion in Behavioral Sciences*, *38*, 116–123. http://dx.doi.org/10.1016/j.cobeha.2021.02.003.

Doya, K., Samejima, K., Katagiri, K., & Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural Computation*, *14*(6), 1347–1369. http://dx.doi.org/10.1162/089976602753712972.

Doya, K., & Uchibe, E. (2005). The cyber rodent project: Exploration of adaptive mechanisms for self-preservation and self-reproduction. *Adaptive Behavior*, *13*(2), 149–160. http://dx.doi.org/10.1177/105971230501300206.

Dunbar, K. (2000). How scientists think in the real world. *Journal of Applied Developmental Psychology*, *21*(1), 49–58. http://dx.doi.org/10.1016/s0193-3973(99)00050-7.

Elfwing, S., & Doya, K. (2014). Emergence of polymorphic mating strategies in robot colonies. *PLoS One*, *9*(4), e93622. http://dx.doi.org/10.1371/journal.pone.0093622.

Elfwing, S., Uchibe, E., Doya, K., & Christensen, H. I. (2011). Darwinian embodied evolution of the learning ability for survival. *Adaptive Behavior*, *19*(2), 101–120. http://dx.doi.org/10.1177/1059712310397633.

Eslami, S. M. A., Jimenez Rezende, D., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., Ruderman, A., Rusu, A. A., Danihelka, I., Gregor, K., Reichert, D. P., Buesing, L., Weber, T., Vinyals, O., Rosenbaum, D., Rabinowitz, N., King, H., Hillier, C., Botvinick, M., ... Hassabis, D. (2018). Neural scene representation and rendering. *Science*, *360*(6394), 1204–1210. http://dx.doi.org/10.1126/science.aar6170.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. http://dx.doi.org/10.1038/nature21056.

Fermin, A. S., Yoshida, T., Yoshimoto, J., Ito, M., Tanaka, S. C., & Doya, K. (2016). Model-based action planning involves cortico-cerebellar and basal ganglia networks. *Scientific Reports*, *6*(31378), http://dx.doi.org/10.1038/srep31378.

Fickinger, A., Zhuang, S., Critch, A., Hadfield-Menell, D., & Russell, S. (2020). Multi-principal assistance games: Definition and collegial mechanisms. In *NeurIPS 2020 workshop on cooperative AI*.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*, 193–202.

Funamizu, A., Kuhn, B., & Doya, K. (2016). Neural substrate of dynamic Bayesian inference in the cerebral cortex. *Nature Neuroscience*, *19*(12), 1682–1689. http://dx.doi.org/10.1038/nn.4390.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411–437. http://dx.doi.org/10.1007/s11023-020-09539-2.

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS Computational Biology*, *11*(11), Article e1004567. http://dx.doi.org/10.1371/journal.pcbi.1004567.

Ghosh, S., Matsuoka, Y., Asai, Y., Hsin, K. Y., & Kitano, H. (2011). Software for systems biology: From tools to integrated platforms. *Nature Reviews Genetics*, *12*(12), 821–832. http://dx.doi.org/10.1038/nrg3096.

Global Neuroethics Summit Delegates, Rommelfanger, K. S., Jeong, S. J., Ema, A., Fukushi, T., Kasai, K., Ramos, K. M., Salles, A., & Singh, I. (2018). Neuroethics questions to guide ethical research in the international brain initiatives. *Neuron*, *100*(1), 19–36. http://dx.doi.org/10.1016/j.neuron.2018.09.021.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Grossberg, S. (2021). *Conscious mind, resonant Brain: How each brain makes a Mind*. Oxford University Press, http://dx.doi.org/10.1093/oso/9780190070557.001.0001.

Ha, D., & Schmidhuber, J. (2018). World models. arXiv:1803.10122.

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The off-switch game. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence* (pp. 220–227). http://dx.doi.org/10.24963/ijcai.2017/32.

Hafner, D., Lillicrap, T., Ba, J., & Norouzi, M. (2020). Dream to control: Learning behaviors by latent imagination. In *ICLR 2020(1912.01603)*.

Hamrick, J. B. (2019). Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences*, *29*, 8–16. http://dx.doi.org/10.1016/j.cobeha.2018.12.011.

Haruno, M., & Frith, C. D. (2010). Activity in the amygdala elicited by unfair divisions predicts social value orientation. *Nature Neuroscience*, *13*(2), 160–161. http://dx.doi.org/10.1038/nn.2468.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258. http://dx.doi.org/10.1016/j.neuron.2017.06.011.

Hawking, S., Tegmark, M., Russell, S., & Wilczek, F. (2014). Transcending complacency on superintelligent machines. Huffpost. April 19, 2014, https://www.huffingtonpost.com/stephen-hawking/artificial-intelligence_b_5174265.html.

Hickok, M. (2020). Lessons learned from AI ethics principles for future actions. *AI and Ethics*, *1*(1), 41–47. http://dx.doi.org/10.1007/s43681-020-00008-1.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal Physiology*, *148*, 574–591. http://dx.doi.org/10.1113/jphysiol.1959.sp006308.

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). *Automated machine learning: Methods, systems, challenges*. Springer, http://dx.doi.org/10.1007/978-3-030-05318-5.

Jamshidi, M., Lalbakhsh, A., Talla, J., Peroutka, Z., Hadjilooei, F., Lalbakhsh, P., Jamshidi, M., Spada, L. L., Mirmozafari, M., Dehghani, M., Sabet, A., Roshani, S., Roshani, S., Bayat-Makou, N., Mohamadzade, B., Malek, Z., Jamshidi, A., Kiani, S., Hashemi-Dezaki, H., & Mohyuddin, W. (2020). Artificial intelligence and COVID-19: Deep learning approaches for diagnosis and treatment. *IEEE Access*, *8*, Article 109581-109595. http://dx.doi.org/10.1109/access.2020.3001973.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. http://dx.doi.org/10.1038/s42256-019-0088-2.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. http://dx.doi.org/10.1038/s41586-021-03819-2.

Jumper, J., & Hassabis, D. (2022). Protein structure predictions to atomic accuracy with AlphaFold. *Nature Methods*, *19*(1), 11–12. http://dx.doi.org/10.1038/s41592-021-01362-6.

Kahneman, D. (2011). *Thinking, fast and Slow*. Farrar, Straus and Giroux.

Kaizu, K., Ghosh, S., Matsuoka, Y., Moriya, H., Shimizu-Yoshida, Y., & Kitano, H. (2010). A comprehensive molecular interaction map of the budding yeast cell cycle. *Molecular Systems Biology*, *6*(415), http://dx.doi.org/10.1038/msb.2010.73.

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L. N., Sparkes, A., Whelan, K. E., & Clare, A. (2009). The automation of science. *Science*, *324*(5923), 85–89. http://dx.doi.org/10.1126/science.1165620.

Kitano, H. (2002). Systems biology: A brief overview. *Science*, *295*(5560), 1662–1664. http://dx.doi.org/10.1126/science.1069492.

Kitano, H. (2016). Artificial intelligence to win the nobel prize and beyond: Creating the engine for scientific discovery. *AI Magazine(Spring 2016)*, 39–49.

Kitano, H. (2021). Nobel turing challenge: Creating the engine for scientific discovery. *Molecular Systems Biology*, *7*(1), 29. http://dx.doi.org/10.1038/s41540-021-00189-3.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*.

Kurth-Nelson, Z., & Redish, A. D. (2009). Temporal-difference reinforcement learning with distributed representations. *PLoS One*, *4*(10), Article e7362. http://dx.doi.org/10.1371/journal.pone.0007362.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, Article e253. http://dx.doi.org/10.1017/S0140525X16001837.

Levine, D. S. (2021). *Healing the reason-emotion split: Scarecrows, tin woodmen, and the wizard*. Routledge.

Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A., & Lederberg, J. (1980). *Applications of artificial intelligence for organic chemistry: The dendral project*. McGraw-Hill.

Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020). Recalibrating global data center energy-use estimates. *Science*, *367*(6481), 984–986. http://dx.doi.org/10.1126/science.aba3758.

Minsky, M., & Papert, S. A. (1969). *Perceptrons: An introduction to computational geometry*. MIT Press, https://mitpress.mit.edu/books/perceptrons.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. http://dx.doi.org/10.1038/s42256-019-0114-4.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. http://dx.doi.org/10.1038/nature14236.

Montes, G. A., & Goertzel, B. (2018). Distributed, decentralized, and democratized artificial intelligence. *Technological Forecasting and Social Change*, http://dx.doi.org/10.1016/j.techfore.2018.11.010.

Morimoto, J., & Doya, K. (2001). Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning. *Robotics and Autonomous Systems*, *36*, 37–51. http://dx.doi.org/10.1016/S0921-8890(01)00113-0.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337. http://dx.doi.org/10.1016/s0896-6273(03)00169-7.

O'Keefe, J. (1976). Place units in the hippocampus of the freely moving rat. *Experimental Neurology*, *51*, 78–109.

Oudeyer, P.-Y. (2018). Computational theories of curiosity-driven learning. In G. Gordon (Ed.), *Psychology of emotions, motivations and actions: The new science of curiosity* (pp. 43–72). Nova Science Publishers, https://arxiv.org/abs/1802.10546.

Palmigiano, A., Geisel, T., Wolf, F., & Battaglia, D. (2017). Flexible information routing by transient synchrony. *Nature Neuroscience*, http://dx.doi.org/10.1038/nn.4569.

Parr, T., & Friston, K. J. (2017). Uncertainty, epistemics and active inference. *Journal of The Royal Society Interface*, *14*(136), http://dx.doi.org/10.1098/rsif.2017.0376.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427–430. http://dx.doi.org/10.1038/nature11467.

Reich, R., Sahami, M., & Weinstein, J. (2021). *System error*. Harper Collins.

Reinke, C. (2018). *The gamma-ensemble: Adaptive reinforcement learning via modular discounting (Publication Number 38005-17)*. Okinawa Institute of Science and Technology Graduate University.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64–99). Appleton-Century-Crofts.

Roach, J. (1997). Austin powers: International man of mystery new line cinema.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., & Waldman-Brown, A. (2019). Tackling climate change with machine learning. arXiv preprint arXiv:1906.05433.

Rosenblatt, F. (1957). *The perceptron — a perceiving and recognizing automaton: Cornell aeronautical laboratory report*, Issue.

Roy, K., Jaiswal, A., & Panda, P. (2019). Towards spike-based machine intelligence with neuromorphic computing. *Nature*, *575*(7784), 607–617. http://dx.doi.org/10.1038/s41586-019-1677-2.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(9), 533–536.

Russell, S. (2015). Take a stand on AI weapons. *Nature*, *521*, 415–416.

Russell, S. (2020). *Human compatible artificial intelligence and the problem of control*. Penguin Random House.

Samejima, K., & Doya, K. (2007). Multiple representations of belief states and action values in corticobasal ganglia loops. *Annals of the New York Academy of Sciences*, *1104*, 213–228. http://dx.doi.org/10.1196/annals.1390.024.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, *310*(5752), 1337–1340. http://dx.doi.org/10.1126/science.1115270.

Schaeffer, H. (2017). Learning partial differential equations via data discovery and sparse optimization. *Proceedings of The Royal Society A Mathematical Physical and Engineering Sciences*, *473*(2197), Article 20160446. http://dx.doi.org/10.1098/rspa.2016.0446.

Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. W W Norton & Co Inc.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., & Silver, D. (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, *588*(7839), 604–609. http://dx.doi.org/10.1038/s41586-020-03051-4.

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, *13*, 900–913.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599. http://dx.doi.org/10.1126/science.275.5306.1593.

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *Elife*, *8*, http://dx.doi.org/10.7554/eLife.41703.

Sejnowski, T. J. (2018). *The deep learning revolution*. MIT Press.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, http://dx.doi.org/10.1038/s41586-019-1923-7.

Shah, R., Freire, P., Alex, N., Freedman, R., Krasheninnikov, D., Chan, L., Dennis, M., Abbeel, P., Dragan, A., & Russell, S. (2020). Benefits of assistance over reward learning. In *NeurIPS 2020 workshop on cooperative AI*.

Shah, R., Krasheninnikov, D., Alexander, J., Abbeel, P., & Dragan, A. (2019). Preferences implicit in the state of the world. In *International conference on learning representations*.

Shirakawa, H., Louis, E. J., MacDiarmid, A. G., Chiang, C. K., & Heeger, A. J. (1977). Synthesis of electrically conducting organic polymers: Halogen derivatives of polyacetylene, (CH) X. *Journal of the Chemical Society, Chemical Communications*, (16), http://dx.doi.org/10.1039/c39770000578.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. http://dx.doi.org/10.1038/nature16961.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354–359. http://dx.doi.org/10.1038/nature24270.

Soto, D., Sheikh, U. A., Mei, N., & Santana, R. (2020). Decoding and encoding models reveal the role of mental simulation in the brain representation of meaning. *Royal Society Open Science*, *7*(5), Article 192043. http://dx.doi.org/10.1098/rsos.192043.

Stray, J. (2020). Aligning AI optimization to community well-being. *International Journal of Community Well-Being*, *3*(4), 443–463. http://dx.doi.org/10.1007/s42413-020-00086-3.

Stray, J., Vendrov, I., Nixon, J., Adler, S., & Hadfield-Menell, D. (2021). What are you optimizing for? aligning recommender systems with human values. arXiv preprint arXiv:2107.10939.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1998). Intra-option learning about temporally abstract actions. In *ICML 1998*.

Takahashi, K., & Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, *126*(4), 663–676. http://dx.doi.org/10.1016/j.cell.2006.07.024.

Taylor, J. (2016). Quantilizers: A safer alternative to maximizers for limited optimization. In *30th AAAI conference on artificial intelligence (AAAI-16) workshop on AI, ethics, and society*. https://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12613.

Thrun, S., & Pratt, L. (Eds.), (1998). *Learning to learn*. Springer, http://dx.doi.org/10.1007/978-1-4615-5529-2.

Turing, A. M. (1950). I.—Computing machinery and intelligence. *Mind*, *LIX*(236), 433–460. http://dx.doi.org/10.1093/mind/LIX.236.433.

Uddin, M., Wang, Y., & Woodbury-Smith, M. (2019). Artificial intelligence for precision medicine in neurodevelopmental disorders. *Npj Digital Medicine*, *2*, 112. http://dx.doi.org/10.1038/s41746-019-0191-0.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860–868. http://dx.doi.org/10.1038/s41593-018-0147-8.

Wang, X. J., & Yang, G. R. (2018). A disinhibitory circuit motif and flexible information routing in the brain. *Current Opinion in Neurobiology*, *49*, 75–83. http://dx.doi.org/10.1016/j.conb.2018.01.002.

Wang, Y., Yin, J., Wang, G., Li, P., Bi, G., Li, S., Xia, X., Song, J., Pei, G., & Zheng, J. C. (2019). Responsibility and sustainability in brain science, technology, and neuroethics in China-A culture-oriented perspective. *Neuron*, *101*(3), 375–379. http://dx.doi.org/10.1016/j.neuron.2019.01.023.

Washington Post (2007). Bots on the ground: In the field of battle (or even above it), robots are a soldier's best friend. May 6, 2007.

Werbos, P. J. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral science*. Harvard University.

Wiener, N. (1960). Some moral and technical consequences of automation. *Science*, *131*(3410), 1355–1358. http://dx.doi.org/10.1126/science.131.3410.1355.

Wiering, M., & Schmidhuber, J. (1998). Hq-learning. *Adaptive Behavior*, *6*, 219–246.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, *11*, 1317–1329.

Xu, J., Yang, P., Xue, S., Sharma, B., Sanchez-Martin, M., Wang, F., Beaty, K. A., Dehan, E., & Parikh, B. (2019). Translating cancer genomics into precision medicine with artificial intelligence: Applications, challenges and future perspectives. *Human Genetics*, *138*(2), 109–124. http://dx.doi.org/10.1007/s00439-019-01970-5.

Yamagishi, T., Takagishi, H., Fermin, A. S., Kanai, R., Li, Y., & Matsumoto, Y. (2016). Cortical thickness of the dorsolateral prefrontal cortex predicts strategic choices in economic games. *Proceedings of the National Academy of Sciences of the United States of America*, http://dx.doi.org/10.1073/pnas.1523940113.

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X. J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, *22*(2), 297–306. http://dx.doi.org/10.1038/s41593-018-0310-2.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692. http://dx.doi.org/10.1016/j.neuron.2005.04.026.

Yuste, R., Goering, S., Arcas, B. A. Y., Bi, G., Carmena, J. M., Carter, A., Fins, J. J., Friesen, P., Gallant, J., Huggins, J. E., Illes, J., Kellmeyer, P., Klein, E., Marblestone, A., Mitchell, C., Parens, E., Pham, M., Rubel, A., Sadato, N., ... Wolpaw, J. (2017). Four ethical priorities for neurotechnologies and AI. *Nature*, *551*(7679), 159–163. http://dx.doi.org/10.1038/551159a.