Erschienen in: Nature ; 568 (2019), 7753. - S. 477-486
https://dx.doi.org/10.1038/s41586-019-1138-y

Machine behaviour

Iyad Rahwan[1,2,3,34]*, Manuel Cebrian[1,34], Nick Obradovich[1,34], Josh Bongard[4], Jean-François Bonnefon[5], Cynthia Breazeal[1], Jacob W. Crandall[6], Nicholas A. Christakis[7,8,9,10], Iain D. Couzin[11,12,13], Matthew O. Jackson[14,15,16], Nicholas R. Jennings[17,18], Ece Kamar[19], Isabel M. Kloumann[20], Hugo Larochelle[21], David Lazer[22,23,24], Richard McElreath[25,26], Alan Mislove[27], David C. Parkes[28,29], Alex 'Sandy' Pentland[1], Margaret E. Roberts[30], Azim Shariff[31], Joshua B. Tenenbaum[32] & Michael Wellman[33]

Machines powered by artificial intelligence increasingly mediate our social, cultural, economic and political interactions. Understanding the behaviour of artificial intelligence systems is essential to our ability to control their actions, reap their benefits and minimize their harms. Here we argue that this necessitates a broad scientific research agenda to study machine behaviour that incorporates and expands upon the discipline of computer science and includes insights from across the sciences. We first outline a set of questions that are fundamental to this emerging field and then explore the technical, legal and institutional constraints on the study of machine behaviour.

In his landmark 1969 book *Sciences of the Artificial*[1], Nobel Laureate Herbert Simon wrote: "Natural science is knowledge about natural objects and phenomena. We ask whether there cannot also be 'artificial' science—knowledge about artificial objects and phenomena." In line with Simon's vision, we describe the emergence of an interdisciplinary field of scientific study. This field is concerned with the scientific study of intelligent machines, not as engineering artefacts, but as a class of actors with particular behavioural patterns and ecology. This field overlaps with, but is distinct from, computer science and robotics. It treats machine behaviour empirically. This is akin to how ethology and behavioural ecology study animal behaviour by integrating physiology and biochemistry—intrinsic properties—with the study of ecology and evolution—properties shaped by the environment. Animal and human behaviours cannot be fully understood without the study of the contexts in which behaviours occur. Machine behaviour similarly cannot be fully understood without the integrated study of algorithms and the social environments in which algorithms operate[2].

At present, the scientists who study the behaviours of these virtual and embodied artificial intelligence (AI) agents are predominantly the same scientists who have created the agents themselves (throughout we use the term 'AI agents' liberally to refer to both complex and simple algorithms used to make decisions). As these scientists create agents to solve particular tasks, they often focus on ensuring the agents fulfil their intended function (although these respective fields are much broader than the specific examples listed here). For example, AI agents should meet a benchmark of accuracy in document classification, facial recognition or visual object detection. Autonomous cars must navigate successfully in a variety of weather conditions; game-playing agents must defeat a variety of human or machine opponents; and data-mining agents must learn which individuals to target in advertising campaigns on social media.

These AI agents have the potential to augment human welfare and well-being in many ways. Indeed, that is typically the vision of their creators. But a broader consideration of the behaviour of AI agents is now critical. AI agents will increasingly integrate into our society and are already involved in a variety of activities, such as credit scoring, algorithmic trading, local policing, parole decisions, driving, online dating and drone warfare[3,4]. Commentators and scholars from diverse fields—including, but not limited to, cognitive systems engineering, human computer interaction, human factors, science, technology and society, and safety engineering—are raising the alarm about the broad, unintended consequences of AI agents that can exhibit behaviours and produce downstream societal effects—both positive and negative—that are unanticipated by their creators[5–8].

In addition to this lack of predictability surrounding the consequences of AI, there is a fear of the potential loss of human oversight over intelligent machines[5] and of the potential harms that are associated with the increasing use of machines for tasks that were once performed directly by humans[9]. At the same time, researchers describe the benefits that AI agents can offer society by supporting and augmenting human decision-making[10,11]. Although discussions of these issues have led to many important insights in many separate fields of academic inquiry[12], with some highlighting safety challenges of autonomous systems[13] and others studying the implications in fairness, accountability and transparency (for example, the ACM conference on fairness, accountability and transparency (https://fatconference.org/)), many questions remain.

This Review frames and surveys the emerging interdisciplinary field of machine behaviour: the scientific study of behaviour exhibited by

[1]Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Institute for Data, Systems & Society, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany. [4]Department of Computer Science, University of Vermont, Burlington, VT, USA. [5]Toulouse School of Economics (TSM-R), CNRS, Université Toulouse Capitole, Toulouse, France. [6]Computer Science Department, Brigham Young University, Provo, UT, USA. [7]Department of Sociology, Yale University, New Haven, CT, USA. [8]Department of Statistics and Data Science, Yale University, New Haven, CT, USA. [9]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA. [10]Yale Institute for Network Science, Yale University, New Haven, CT, USA. [11]Department of Collective Behaviour, Max Planck Institute for Ornithology, Konstanz, Germany. [12]Department of Biology, University of Konstanz, Konstanz, Germany. [13]Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Konstanz, Germany. [14]Department of Economics, Stanford University, Stanford, CA, USA. [15]Canadian Institute for Advanced Research, Toronto, Ontario, Canada. [16]The Sante Fe Institute, Santa Fe, NM, USA. [17]Department of Computing, Imperial College London, London, UK. [18]Department of Electrical and Electronic Engineering, Imperial College London, London, UK. [19]Microsoft Research, Redmond, WA, USA. [20]Facebook AI, Facebook Inc, New York, NY, USA. [21]Google Brain, Montreal, Québec, Canada. [22]Department of Political Science, Northeastern University, Boston, MA, USA. [23]College of Computer & Information Science, Northeastern University, Boston, MA, USA. [24]Institute for Quantitative Social Science, Harvard University, Cambridge, MA, USA. [25]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [26]Department of Anthropology, University of California, Davis, Davis, CA, USA. [27]College of Computer & Information Science, Northeastern University, Boston, MA, USA. [28]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. [29]Harvard Data Science Initiative, Harvard University, Cambridge, MA, USA. [30]Department of Political Science, University of California, San Diego, San Diego, CA, USA. [31]Department of Psychology, University of British Columbia, Vancouver, British Columbia, Canada. [32]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [33]Computer Science & Engineering, University of Michigan, Ann Arbor, MI, USA. [34]These authors contributed equally: Iyad Rahwan, Manuel Cebrian, Nick Obradovich.
*e-mail: irahwan@mit.edu

intelligent machines. Here we outline the key research themes, questions and landmark research studies that exemplify this field. We start by providing background on the study of machine behaviour and the necessarily interdisciplinary nature of this science. We then provide a framework for the conceptualization of studies of machine behaviour. We close with a call for the scientific study of machine and human–machine ecologies and discuss some of the technical, legal and institutional barriers that are faced by researchers in this field.

Motivation for the study of machine behaviour

There are three primary motivations for the scientific discipline of machine behaviour. First, various kinds of algorithms operate in our society, and algorithms have an ever-increasing role in our daily activities. Second, because of the complex properties of these algorithms and the environments in which they operate, some of their attributes and behaviours can be difficult or impossible to formalize analytically. Third, because of their ubiquity and complexity, predicting the effects of intelligent algorithms on humanity—whether positive or negative—poses a substantial challenge.

Ubiquity of algorithms

The current prevalence of diverse algorithms in society is unprecedented[5] (Fig. 1). News-ranking algorithms and social media bots influence the information seen by citizens[14–18]. Credit-scoring algorithms determine loan decisions[19–22]. Online pricing algorithms shape the cost of products differentially across consumers[23–25]. Algorithmic trading software makes transactions in financial markets at rapid speed[26–29]. Algorithms shape the dispatch and spatial patterns of local policing[30] and programs for algorithmic sentencing affect time served in the penal system[7]. Autonomous cars traverse our cities[31], and ride-sharing algorithms alter the travel patterns of conventional vehicles[32]. Machines map our homes, respond to verbal commands[33] and perform regular household tasks[34]. Algorithms shape romantic matches for online dating services[35,36]. Machines are likely to increasingly substitute for humans in the raising of our young[37] and the care for our old[38]. Autonomous agents are increasingly likely to affect collective behaviours, from group-wide coordination to sharing[39]. Furthermore, although the prospect of developing autonomous weapons is highly controversial, with many in the field voicing their opposition[6,40], if such weapons end up being deployed, then machines could determine who lives and who dies in armed conflicts[41,42].

Complexity and opacity of algorithms

The extreme diversity of these AI systems, coupled with their ubiquity, would by itself ensure that studying the behaviour of such systems poses a formidable challenge, even if the individual algorithms themselves were relatively simple. The complexity of individual AI agents is currently high and rapidly increasing. Although the code for specifying the architecture and training of a model can be simple, the results can be very complex, oftentimes effectively resulting in 'black boxes'[43]. They are given input and produce output, but the exact functional processes that generate these outputs are hard to interpret even to the very scientists who generate the algorithms themselves[44], although some progress in interpretability is being made[45,46]. Furthermore, when systems learn from data, their failures are linked to imperfections in the data or how data was collected, which has led some to argue for adapted reporting mechanisms for datasets[47] and models[48]. The dimensionality and size of data add another layer of complexity to understanding machine behaviour[49].

Further complicating this challenge is the fact that much of the source code and model structure for the most frequently used algorithms in society is proprietary, as are the data on which these systems are trained. Industrial secrecy and legal protection of intellectual property often surround source code and model structure. In many settings, the only factors that are publicly observable about industrial AI systems are their inputs and outputs.

Even when available, the source code or model structure of an AI agent can provide insufficient predictive power over its output. AI agents can also demonstrate novel behaviours through their interaction with the world and other agents that are impossible to predict with precision[50]. Even when the analytical solutions are mathematically describable, they can be so lengthy and complex as to be indecipherable[51,52]. Furthermore, when the environment is changing—perhaps as a result of the algorithm itself—anticipating and analysing behaviour is made much harder.

Algorithms' beneficial and detrimental effect on humanity

The ubiquity of algorithms, coupled with their increasing complexity, tends to amplify the difficulty of estimating the effects of algorithms on individuals and society. AI agents can shape human behaviours and societal outcomes in both intended and unintended ways. For example, some AI agents are designed to aid learning outcomes for children[53] and others are designed to assist older people[38,54]. These AI systems may benefit their intended humans by nudging those humans into better learning or safer mobility behaviours. However, with the power to nudge human behaviours in positive or intended ways comes the risk that human behaviours may be nudged in costly or unintended ways—children could be influenced to buy certain branded products and elders could be nudged to watch certain television programs.

The way that such algorithmic influences on individual humans scale into society-wide effects, both positive and negative, is of critical concern. As an example, the exposure of a small number of individuals to political misinformation may have little effect on society as a whole. However, the effect of the insertion and propagation of such misinformation on social media may have more substantial societal consequences[55–57]. Furthermore, issues of algorithmic fairness or bias[58,59] have been already documented in diverse contexts, including computer vision[60], word embeddings[61,62], advertising[63], policing[64], criminal justice[7,65] and social services[66]. To address these issues, practitioners will sometimes be forced to make value trade-offs between competing and incompatible notions of bias[58,59] or between human versus machine biases. Additional questions regarding the effect of algorithms remain, such as how online dating algorithms alter the societal institution of marriage[35,36] and whether there are systemic effects of increasing interaction with intelligent algorithms on the stages and speed of human development[53]. These questions become more complex in 'hybrid systems' composed of many machines and humans interacting and manifesting collective behaviour[39,67]. For society to have input into and oversight of the downstream consequences of AI, scholars of machine behaviour must provide insights into how these systems work and the benefits, costs and trade-offs presented by the ubiquitous use of AI in society.

The interdisciplinary study of machine behaviour

To study machine behaviour—especially the behaviours of black box algorithms in real-world settings—we must integrate knowledge from across a variety of scientific disciplines (Fig. 2). This integration is currently in its nascent stages and has happened largely in an ad hoc fashion in response to the growing need to understand machine behaviour. Currently, the scientists who most commonly study the behaviour of machines are the computer scientists, roboticists and engineers who have created the machines in the first place. These scientists may be expert mathematicians and engineers; however, they are typically not trained behaviourists. They rarely receive formal instruction on experimental methodology, population-based statistics and sampling paradigms, or observational causal inference, let alone neuroscience, collective behaviour or social theory. Conversely, although behavioural scientists are more likely to possess training in these scientific methods, they are less likely to possess the expertise required to proficiently evaluate the underlying quality and appropriateness of AI techniques for a given problem domain or to mathematically describe the properties of particular algorithms.

Integrating scientific practices from across multiple fields is not easy. Up to this point, the main focus of those who create AI systems has been on crafting, implementing and optimizing intelligent systems to

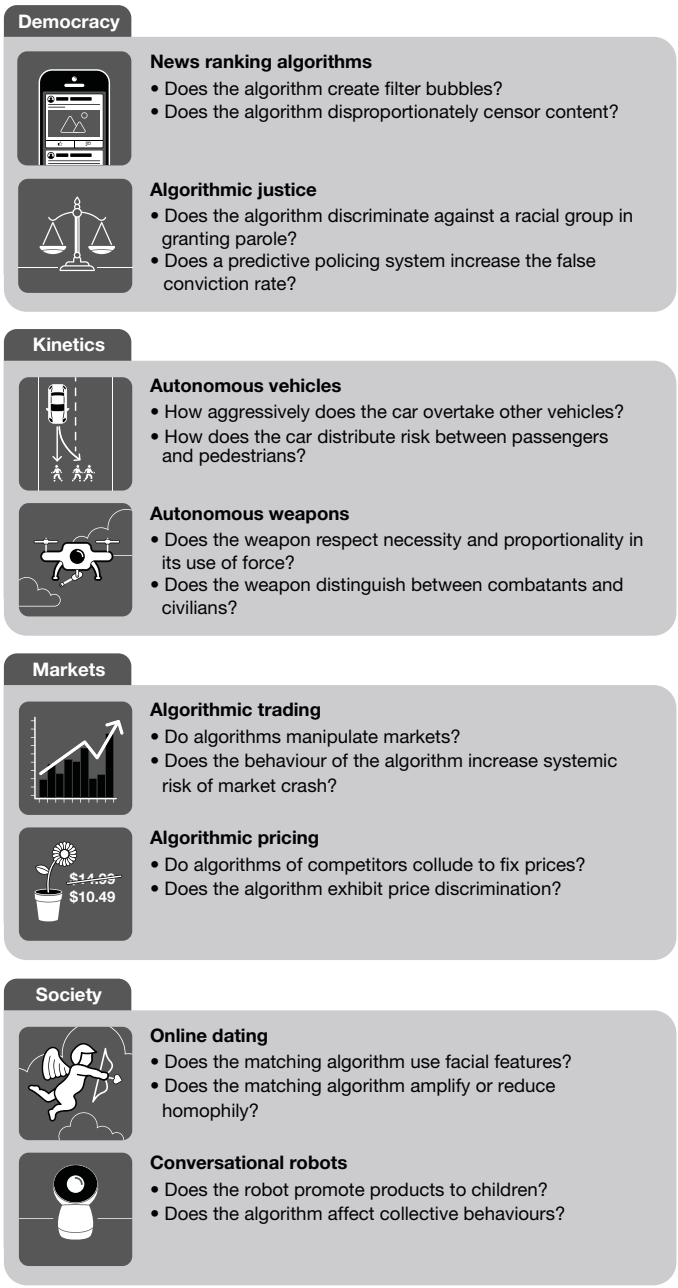

Fig. 1 | Examples of questions that fall into the domain of machine behaviour. Questions of concern to machine behaviour span a wide variety of traditional scientific disciplines and topics.

perform specialized tasks. Excellent progress has been made on benchmark tasks—including board games such as chess[68], checkers[69] and Go[70,71], card games such as poker[72], computer games such as those on the Atari platform[73], artificial markets[74] and Robocup Soccer[75]—as well as standardized evaluation data, such as the ImageNet data for object recognition[76] and the Microsoft Common Objects in Context data for image-captioning tasks[77]. Success has also been achieved in speech recognition, language translation and autonomous locomotion. These benchmarks are coupled with metrics to quantify performance on standardized tasks[78–81] and are used to improved performance, a proxy that enables AI builders to aim for better, faster and more-robust algorithms.

But methodologies aimed at maximized algorithmic performance are not optimal for conducting scientific observation of the properties and behaviours of AI agents. Rather than using metrics in the service of optimization against benchmarks, scholars of machine behaviour are interested in a broader set of indicators, much as social scientists explore a wide range of human behaviours in the realm of social, political or economic interactions[82]. As such, scholars of machine behaviour spend considerable effort in defining measures of micro and macro outcomes to answer broad questions such as how these algorithms behave in different environments and whether human interactions with algorithms alter societal outcomes. Randomized experiments, observational inference and population-based descriptive statistics—methods that are often used in quantitative behavioural sciences—must be central to the study of machine behaviour. Incorporating scholars from outside of the disciplines that traditionally produce intelligent machines can provide knowledge of important methodological tools, scientific approaches, alternative conceptual frameworks and perspectives on the economic, social and political phenomena that machines will increasingly influence.

Type of question and object of study

Nikolaas Tinbergen, who won the 1973 Nobel Prize in Physiology or Medicine alongside Karl von Frisch and Konrad Lorenz for founding the field of ethology, identified four complementary dimensions of

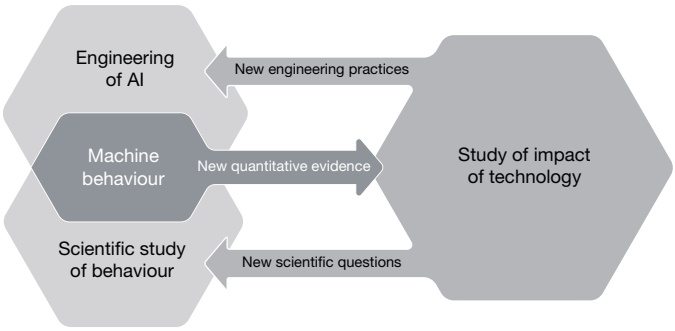

Fig. 2 | The interdisciplinarity of machine behaviour. Machine behaviour lies at the intersection of the fields that design and engineer AI systems and the fields that traditionally use scientific methods to study the behaviour of biological agents. The insights from machine behavioural studies provide quantitative evidence that can help to inform those fields that study the potential effects of technology on social and technological systems. In turn, those fields can provide useful engineering practices and scientific questions to fields that examine machine behaviours. Finally, the scientific study of behaviour helps AI scholars to make more precise statements about what AI systems can and cannot do.

analysis that help to explain animal behaviour[83]. These dimensions concern questions of the function, mechanism, development and evolutionary history of a behaviour and provide an organizing framework for the study of animal and human behaviour. For example, this conceptualization distinguishes the study of how a young animal or human develops a type of behaviour from the evolutionary trajectory that selected for such behaviour in the population. The goal of these distinctions is not division but rather integration. Although it is not wrong to say that, for example, a bird's song is explained by learning or by its specific evolutionary history, a complete understanding of the song will require both.

Despite fundamental differences between machines and animals, the behavioural study of machines can benefit from a similar classification. Machines have mechanisms that produce behaviour, undergo development that integrates environmental information into behaviour, produce functional consequences that cause specific machines to become more or less common in specific environments and embody evolutionary histories through which past environments and human decisions continue to influence machine behaviour. Scholars of computer science have already achieved substantial gains in understanding the mechanisms and development of AI systems, although many questions remain. Relatively less emphasis has been placed on the function and evolution of AI systems. We discuss these four topics in the next subsections and provide Fig. 3 as a summary[84].

Mechanisms for generating behaviour

The proximate causes of a machine's behaviour have to do with how the behaviour is observationally triggered and generated in specific environments. For example, early algorithmic trading programs used simple rules to trigger buying and selling behaviour[85]. More sophisticated agents may compute strategies based on adaptive heuristics or explicit maximization of expected utility[86]. The behaviour of a reinforcement learning algorithm that plays poker could be attributed to the particular way in which it represents the state space or evaluates the game tree[72], and so on.

A mechanism depends on both an algorithm and its environment. A more sophisticated agent, such as a driverless car, may exhibit particular driving behaviour—for example, lane switching, overtaking or signalling to pedestrians. These behaviours would be generated according to the algorithms that construct driving policies[87] and are also shaped fundamentally by features of the perception and actuation system of the car, including the resolution and accuracy of its object detection and classification system, and the responsiveness and accuracy of its steering, among other factors. Because many current AI systems are derived from machine learning methods that are applied to increasingly complex data, the study of the mechanism behind a machine's behaviour, such as those mentioned above, will require continued work on interpretability methods for machine learning[46,88,89].

Development of behaviour

In the study of animal or human behaviour, development refers to how an individual acquires a particular behaviour—for example, through imitation or environmental conditioning. This is distinct from longer-term evolutionary changes.

In the context of machines, we can ask how machines acquire (develop) a specific individual or collective behaviour. Behavioural development could be directly attributable to human engineering or design choices. Architectural design choices made by the programmer (for example, the value of a learning rate parameter, the acquisition of the representation of knowledge and state, or a particular wiring of a convolutional neural network) determine or influence the kinds of behaviours that the algorithm exhibits. In a more complex AI system, such as a driverless car, the behaviour of the car develops over time, from software development and changing hardware components that engineers incorporate into its overall architecture. Behaviours can also change as a result of algorithmic upgrades pushed to the machine by its designers after deployment.

A human engineer may also shape the behaviour of the machine by exposing it to particular training stimuli. For instance, many image and text classification algorithms are trained to optimize accuracy on a specific set of datasets that were manually labelled by humans. The choice of dataset—and those features it represents[60,61]—can substantially influence the behaviour exhibited by the algorithm.

Finally, a machine may acquire behaviours through its own experience. For instance, a reinforcement learning agent trained to maximize long-term profit can learn peculiar short-term trading strategies based on its own past actions and concomitant feedback from the market[90]. Similarly, product recommendation algorithms make recommendations based on an endless stream of choices made by customers and update their recommendations accordingly.

Function

In the study of animal behaviour, adaptive value describes how a behaviour contributes to the lifetime reproductive fitness of an animal. For example, a particular hunting behaviour may be more or less successful than another at prolonging the animal's life and, relatedly, the number of mating opportunities, resulting offspring born and the probable reproductive success of the offspring. The focus on function helps us to understand why some behavioural mechanisms spread and persist while others decline and vanish. Function depends critically on the fit of the behaviour to environment.

In the case of machines, we may talk of how the behaviour fulfils a contemporaneous function for particular human stakeholders. The human environment creates selective forces that may make some machines more common. Behaviours that are successful ('fitness' enhancing) get copied by developers of other software and hardware or are sometimes engineered to propagate among the machines themselves. These dynamics are ultimately driven by the success of institutions—such as corporations, hospitals, municipal governments and universities—that build or use AI. The most obvious example is provided by algorithmic trading, in which successful automated trading strategies could be copied as their developers move from company to company, or are simply observed and reverse-engineered by rivals.

These forces can produce unanticipated effects. For example, objectives such as maximizing engagement on a social media site may lead to so-called filter bubbles[91], which may increase political polarization or, without careful moderation, could facilitate the spread of fake news. However, websites that do not optimize for user engagement may not be as successful in comparison with ones that do, or may go out of business altogether. Similarly, in the absence of external regulation, autonomous cars that do not prioritize the safety of their own passengers may be

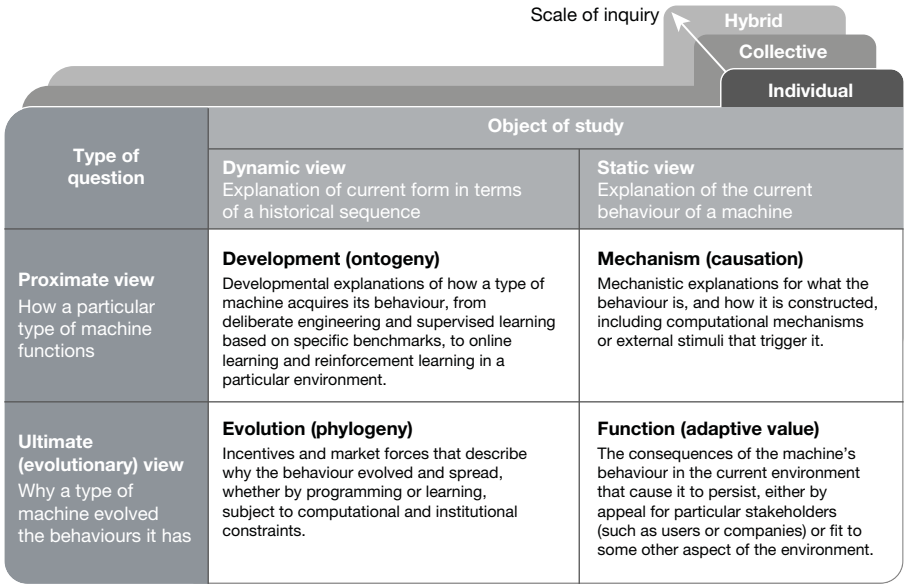

Fig. 3 | Tinbergen's type of question and object of study modified for the study of machine behaviour. The four categories Tinbergen proposed for the study of animal behaviour can be adapted to the study of machine behaviour[83,84]. Tinbergen's framework proposes two types of question, how versus why, as well as two views of these questions, dynamic versus static. Each question can be examined at three scales of inquiry: individual machines, collectives of machines and hybrid human–machine systems.

less attractive to consumers, leading to fewer sales[31]. Sometimes the function of machine behaviour is to cope with the behaviour of other machines. Adversarial attacks—synthetic inputs that fool a system into producing an undesired output[44,92–94]—on AI systems and the subsequent responses of those who develop AI to these attacks[95] may produce complex predator–prey dynamics that are not easily understood by studying each machine in isolation.

These examples highlight how incentives created by external institutions and economic forces can have indirect but substantial effects on the behaviours exhibited by machines[96]. Understanding the interaction between these incentives and AI is relevant to the study of machine behaviour. These market dynamics would, in turn, interact with other processes to produce evolution among machines and algorithms.

Evolution

In the study of animal behaviour, phylogeny describes how a behaviour evolved. In addition to its current function, behaviour is influenced by past selective pressures and previously evolved mechanisms. For example, the human hand evolved from the fin of a bony fish. Its current function is no longer for swimming, but its internal structure is explained by its evolutionary history. Non-selective forces, such as migration and drift, also have strong roles in explaining relationships among different forms of behaviour.

In the case of machines, evolutionary history can also generate path dependence, explaining otherwise puzzling behaviour. At each step, aspects of the algorithms are reused in new contexts, both constraining future behaviour and making possible additional innovations. For example, early choices about microprocessor design continue to influence modern computing, and traditions in algorithm design—such as neural networks and Bayesian state–space models—build in many assumptions and guide future innovations by making some new algorithms easier to access than others. As a result, some algorithms may attend to certain features and ignore others because those features were important in early successful applications. Some machine behaviour may spread because it is 'evolvable'—easy to modify and robust to perturbations—similar to how some traits of animals may be common because they facilitate diversity and stability[97].

Machine behaviour evolves differently from animal behaviour. Most animal inheritance is simple—two parents, one transmission event. Algorithms are much more flexible and they have a designer with an objective in the background. The human environment strongly influences how algorithms evolve by changing their inheritance system. AI replication behaviour may be facilitated through a culture of open source sharing of software, the details of network architecture or underlying training datasets. For instance, companies that develop software for driverless cars may share enhanced open source libraries for object detection or path planning as well as the training data that underlie these algorithms to enable safety-enhancing software to spread throughout the industry. It is possible for a single adaptive 'mutation' in the behaviour of a particular driverless car to propagate instantly to millions of other cars through a software update. However, other institutions apply limits as well. For example, software patents may impose constraints on the copying of particular behavioural traits. And regulatory constraints—such as privacy protection laws—can prevent machines from accessing, retaining or otherwise using particular information in their decision-making. These peculiarities highlight the fact that machines may exhibit very different evolutionary trajectories, as they are not bound by the mechanisms of organic evolution.

Scale of inquiry

With the framework outlined above and in Fig. 3, we now catalogue examples of machine behaviour at the three scales of inquiry: individual machines, collectives of machines and groups of machines embedded in a social environment with groups of humans in hybrid or heterogeneous systems[39] (Fig. 4). Individual machine behaviour emphasizes the study of the algorithm itself, collective machine behaviour emphasizes the study of interactions between machines and hybrid human–machine behaviour emphasizes the study of interactions between machines and humans. Here we can draw an analogy to the study of a particular species, the study of interactions among members of a species and the interactions of the species with their broader environment. Analyses at any of these scales may address any or all of the questions described in Fig. 3.

Individual machine behaviour

The study of the behaviour of individual machines focuses on specific intelligent machines by themselves. Often these studies focus on properties that are intrinsic to the individual machines and that are driven by their source code or design. The fields of machine learning and software engineering currently conduct the majority of these studies. There are two general approaches to the study of individual machine behaviour. The first focuses on profiling the set of behaviours of any

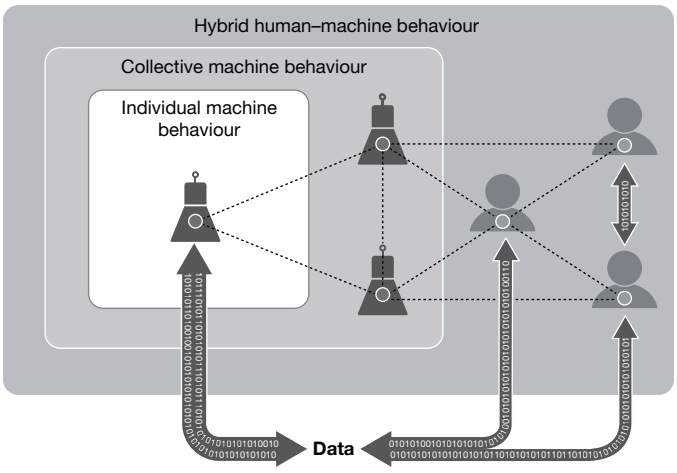

Fig. 4 | Scale of inquiry in the machine behaviour ecosystem. AI systems represent the amalgamation of humans, data and algorithms. Each of these domains influences the other in both well-understood and unknown ways. Data—filtered through algorithms created by humans—influences individual and collective machine behaviour. AI systems are trained on the data, in turn influencing how humans generate data. AI systems collectively interact with and influence one another. Human interactions can be altered by the introduction of these AI systems. Studies of machine behaviour tend to occur at the individual, the collective or the hybrid human–machine scale of inquiry.

specific machine agent using a within-machine approach, comparing the behaviour of a particular machine across different conditions. The second, a between-machine approach, examines how a variety of individual machine agents behave in the same condition.

A within-machine approach to the study of individual machine behaviours investigates questions such as whether there are constants that characterize the within-machine behaviour of any particular AI across a variety of contexts, how the behaviour of a particular AI progresses over time in the same, or different, environments and which environmental factors lead to the expression of particular behaviours by machines.

For instance, an algorithm may only exhibit certain behaviours if trained on particular underlying data[98–100] (Fig. 3). Then, the question becomes whether or not an algorithm that scores probability of recidivism in parole decisions[7] would behave in unexpected ways when presented with evaluation data that diverge substantially from its training data. Other studies related to the characterization of within-machine behaviour include the study of individual robotic recovery behaviours[101,102], the 'cognitive' attributes of algorithms and the utility of using techniques from psychology in the study of algorithmic behaviour[103], and the examination of bot-specific characteristics such as those designed to influence human users[104].

The second approach to the study of individual machine behaviour examines the same behaviours as they vary between machines. For example, those interested in examining advertising behaviours of intelligent agents[63,105,106] may investigate a variety of advertising platforms (and their underlying algorithms) and examine the between-machine effect of performing experiments with the same set of advertising inputs across platforms. The same approach could be used for investigations of dynamic pricing algorithms[23,24,32] across platforms. Other between-machine studies might look at the different behaviours used by autonomous vehicles in their overtaking patterns or at the varied foraging behaviours exhibited by search and rescue drones[107].

Collective machine behaviour

In contrast the study of the behaviour of individual machines, the study of collective machine behaviour focuses on the interactive and system-wide behaviours of collections of machine agents. In some cases, the implications of individual machine behaviour may make little sense until the collective level is considered. Some investigations of these systems have been inspired by natural collectives, such as swarms of insects, or mobile groups, such as flocking birds or schooling fish. For example, animal groups are known to exhibit both emergent sensing of complex environmental features[108] and effective consensus decision-making[109]. In both scenarios, groups exhibit an awareness of the environment that does not exist at the individual level. Fields such as multi-agent systems and computational game theory provide useful examples of the study of this area of machine behaviour.

Robots that use simple algorithms for local interactions between bots can nevertheless produce interesting behaviour once aggregated into large collectives. For example, scholars have examined the swarm-like properties of microrobots that combine into aggregations that resemble swarms found in systems of biological agents[110,111]. Additional examples include the collective behaviours of algorithms both in the laboratory (in the Game of Life[112]) as well as in the wild (as seen in Wikipedia-editing bots[113]). Other examples include the emergence of novel algorithmic languages[114] between communicating intelligent machines as well as the dynamic properties of fully autonomous transportation systems. Ultimately, many interesting questions in this domain remain to be examined.

The vast majority of work on collective animal behaviour and collective robotics has focused on how interactions among simple agents can create higher-order structures and properties. Although important, this neglects that fact that many organisms, and increasingly also AI agents[75], are sophisticated entities with behaviours and interactions that may not be well-characterized by simplistic representations. Revealing what extra properties emerge when interacting entities are capable of sophisticated cognition remains a key challenge in the biological sciences and may have direct parallels in the study of machine behaviour. For example, similar to animals, machines may exhibit 'social learning'. Such social learning does not need be limited to machines learning from machines, but we may expect machines to learn from humans, and vice versa for humans to learn from the behaviour of machines. The feedback processes introduced may fundamentally alter the accumulation of knowledge, including across generations, directly affecting human and machine 'culture'.

In addition, human-made AI systems do not necessarily face the same constraints as do organisms, and collective assemblages of machines provide new capabilities, such as instant global communication, that can lead to entirely new collective behavioural patterns. Studies in collective machine behaviour examine the properties of assemblages of machines as well as the unexpected properties that can emerge from these complex systems of interactions.

For example, some of the most interesting collective behaviour of algorithms has been observed in financial trading environments. These environments operate on tiny time scales, such that algorithmic traders can respond to events and each other ahead of any human trader[115]. Under certain conditions, high-frequency capabilities can produce inefficiencies in financial markets[26,115]. In addition to the unprecedented response speed, the extensive use of machine learning, autonomous operation and ability to deploy at scale are all reasons to believe that the collective behaviour of machine trading may be qualitatively different than that of human traders. Furthermore, these financial algorithms and trading systems are necessarily trained on certain historic datasets and react to a limited variety of foreseen scenarios, leading to the question of how they will react to situations that are new and unforeseen in their design. Flash crashes are examples of clearly unintended consequences of (interacting) algorithms[116,117]; leading to the question of whether algorithms could interact to create a larger market crisis.

Hybrid human–machine behaviour

Humans increasingly interact with machines[16]. They mediate our social interactions[39], shape the news[14,17,55,56] and online information[15,118] that we see, and form relationships with us that can alter our social systems. Because of their complexity, these hybrid human–machine systems

pose one of the most technically difficult yet simultaneously most important areas of study for machine behaviour.

Machines shape human behaviour

One of the most obvious—but nonetheless vital—domains of the study of machine behaviour concerns the ways in which the introduction of intelligent machines into social systems can alter human beliefs and behaviours. As in the introduction of automation to industrial processes[119], intelligent machines can create social problems in the process of improving existing problems. Numerous problems and questions arise during this process, such as whether the matching algorithms that are used for online dating alter the distributional outcomes of the dating process or whether news-filtering algorithms alter the distribution of public opinion. It is important to investigate whether small errors in algorithms or the data that they use could compound to produce society-wide effects and how intelligent robots in our schools, hospitals[120] and care centres might alter human development[121] and quality of life[54] and potentially affect outcomes for people with disabilities[122].

Other questions in this domain relate to the potential for machines to alter the social fabric in more fundamental ways. For example, questions include to what extent and what ways are governments using machine intelligence to alter the nature of democracy, political accountability and transparency, or civic participation. Other questions include to what degree intelligent machines influence policing, surveillance and warfare, as well as how large of an effect bots have had on the outcomes of elections[56] and whether AI systems that aid in the formation of human social relationships can enable collective action.

Notably, studies in this area also examine how humans perceive the use of machines as decision aids[8,123], human preferences for and against making use of algorithms[124], and the degree to which human-like machines produce or reduce discomfort in humans[39,125]. An important question in this area includes how humans respond to the increasing coproduction of economic goods and services in tandem with intelligent machines[126]. Ultimately, understanding how human systems can be altered by the introduction of intelligent machines into our lives is a vital component of the study of machine behaviour.

Humans shape machine behaviour

Intelligent machines can alter human behaviour, and humans also create, inform and mould the behaviours of intelligent machines. We shape machine behaviours through the direct engineering of AI systems and through the training of these systems on both active human input and passive observations of human behaviours through the data that we create daily. The choice of which algorithms to use, what feedback to provide to those algorithms[3,127] and on which data to train them are also, at present, human decisions and can directly alter machine behaviours. An important component in the study of machine behaviour is to understand how these engineering processes alter the resulting behaviours of AI, whether the training data are responsible for a particular behaviour of the machine, whether it is the algorithm itself or whether it is a combination of both algorithm and data. The framework outlined in Fig. 3 suggests that there will be complementary answers to the each of these questions. Examining how altering the parameters of the engineering process can alter the subsequent behaviours of intelligent machines as they interact with other machines and with humans in natural settings is central to a holistic understanding of machine behaviour.

Human–machine co-behaviour

Although it can be methodologically convenient to separate studies into the ways that humans shape machines and vice versa, most AI systems function in domains where they co-exist with humans in complex hybrid systems[39,67,125,128]. Questions of importance to the study of these systems include those that examine the behaviours that characterize human–machine interactions including cooperation, competition and coordination—for example, how human biases combine with AI to alter human emotions or beliefs[14,55,56,129,130], how human tendencies couple with algorithms to facilitate the spread of information[55], how traffic patterns can be altered in streets populated by large numbers of both driverless and human-driven cars and how trading patterns can be altered by interactions between humans and algorithmic trading agents[29] as well as which factors can facilitate trust and cooperation between humans and machines[88,131].

Another topic in this area relates to robotic and software-driven automation of human labour[132]. Here we see two different types of machine–human interactions. One is that machines can enhance a human's efficiency, such as in robotic- and computer-aided surgery. Another is that machines can replace humans, such as in driverless transportation and package delivery. This leads to questions about whether machines end up doing more of the replacing or the enhancing in the longer run and what human–machine co-behaviours will evolve as a result.

The above examples highlight that many of the questions that relate to hybrid human–machine behaviours must necessarily examine the feedback loops between human influence on machine behaviour and machine influence on human behaviour simultaneously. Scholars have begun to examine human–machine interactions in formal laboratory environments, observing that interactions with simple bots can increase human coordination[39] and that bots can cooperate directly with humans at levels that rival human–human cooperation[133]. However, there remains an urgent need to further understand feedback loops in natural settings, in which humans are increasingly using algorithms to make decisions[134] and subsequently informing the training of the same algorithms through those decisions. Furthermore, across all types of questions in the domain of machine behavioural ecology, there is a need for studies that examine longer-run dynamics of these hybrid systems[53] with particular emphasis on the ways that human social interactions[135,136] may be modified by the introduction of intelligent machines[137].

Outlook

Furthering the study of machine behaviour is critical to maximizing the potential benefits of AI for society. The consequential choices that we make regarding the integration of AI agents into human lives must be made with some understanding of the eventual societal implications of these choices. To provide this understanding and anticipation, we need a new interdisciplinary field of scientific study: machine behaviour.

For this field to succeed, there are a number of relevant considerations. First, studying machine behaviour does not imply that AI algorithms necessarily have independent agency nor does it imply algorithms should bear moral responsibility for their actions. If a dog bites someone, the dog's owner is held responsible. Nonetheless, it is useful to study the behavioural patterns of animals to predict such aberrant behaviour. Machines operate within a larger socio-technical fabric, and their human stakeholders are ultimately responsible for any harm their deployment might cause.

Second, some commentators might suggest that treating AI systems as agents occludes the focus on the underlying data that such AI systems are trained on. Indeed, no behaviour is ever fully separable from the environmental data on which that agent is trained or developed; machine behaviour is no exception. However, it is just as critical to understand how machine behaviours vary with altered environmental inputs as it is to understand how biological agents' behaviours vary depending on the environments in which they exist. As such, scholars of machine behaviour should focus on characterizing agent behaviour across diverse environments, much as behavioural scientists desire to characterize political behaviours across differing demographic and institutional contexts.

Third, machines exhibit behaviours that are fundamentally different from animals and humans, so we must avoid excessive anthropomorphism and zoomorphism. Even if borrowing existing behavioural scientific methods can prove useful for the study of machines, machines may exhibit forms of intelligence and behaviour that are qualitatively different—even alien—from those seen in biological agents.

Furthermore, AI scientists can dissect and modify AI systems more easily and more thoroughly than is the case for many living systems. Although parallels exist, the study of AI systems will necessarily differ from the study of living systems.

Fourth, the study of machine behaviour will require cross-disciplinary efforts[82,103] and will entail all of the challenges associated with such research[138,139]. Addressing these challenges is vital[140]. Universities and governmental funding agencies can play an important part in the design of large-scale, neutral and trusted cross-disciplinary studies[141].

Fifth, the study of machine behaviour will often require experimental intervention to study human–machine interactions in real-world settings[142,143]. These interventions could alter the overall behaviour of the system, possibly having adverse effects on normal users[144]. Ethical considerations such as these need careful oversight and standardized frameworks.

Finally, studying intelligent algorithmic or robotic systems can result in legal and ethical problems for researchers studying machine behaviour. Reverse-engineering algorithms may require violating the terms of service of some platforms; for example, in setting up fake personas or masking true identities. The creators or maintainers of the systems of interest could embroil researchers in legal challenges if the research damages the reputation of their platforms. Moreover, it remains unclear whether violating terms of service may expose researchers to civil or criminal penalties (for example, through the Computer Fraud and Abuse Act in the United States), which may further discourage this type of research[145].

Understanding the behaviours and properties of AI agents—and the effects they might have on human systems—is critical. Society can benefit tremendously from the efficiencies and improved decision-making that can come from these agents. At the same time, these benefits may falter without minimizing the potential pitfalls of the incorporation of AI agents into everyday human life.

1. Simon, H. A. *The Sciences of the Artificial* (MIT Press, Cambridge, 1969). **Simon asks whether there can be a science of the 'artificial' that produces knowledge about artificial objects and phenomena**.
2. Milner, R. A modal characterisation of observable machine-behaviour. In *Trees in Algebra and Programming, 6th Colloquium* 25–34 (Springer, 1981). **In this invited lecture, Robin Milner outlines the idea of studying machine behaviour using formal logic**.
3. Thomaz, A. L. & Breazeal, C. Teachable robots: understanding human teaching behavior to build more effective robot learners. *Artif. Intell.* **172**, 716–737 (2008).
4. Stone, P. *et al. Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015–2016 Study Panel* https://ai100.stanford.edu/2016-report (Stanford University, 2016).
5. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Broadway Books, 2016). **This book articulates some of the risks posed by the uncritical use of algorithms in society and provides motivation for the study of machine behaviour**.
6. Future of Life Institute. Autonomous weapons: an open letter from AI & robotics researchers. https://futureoflife.org/open-letter-autonomous-weapons/?cn-reloaded=1 (2015).
7. Dressel, J. & Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* **4**, eaao5580 (2018).
8. Binns, R. *et al.* 'It's reducing a human being to a percentage': perceptions of justice in algorithmic decisions. In *Proc. 2018 CHI Conference on Human Factors in Computing Systems* 377 (ACM, 2018).
9. Hudson, L., Owens, C. S. & Flannes, M. Drone warfare: blowback from the new American way of war. *Middle East Policy* **18**, 122–132 (2011).
10. Kahneman, D., Rosenfield, A. M., Gandhi, L. & Blaser, T. Noise: how to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review* https://hbr.org/2016/10/noise (2016).
11. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. & Mullainathan, S. Human decisions and machine predictions. *Q. J. Econ.* **133**, 237–293 (2018).
12. Crawford, K. et al. *The AI Now report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-term*. https://ainowinstitute.org/AI_Now_2016_Report.pdf (2016).
13. Amodei, D. et al. Concrete problems in AI safety. Preprint at https://arxiv.org/abs/1606.06565 (2016).
14. Bakshy, E., Messing, S. & Adamic, L. A. Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132 (2015).
15. Bessi, A. & Ferrara, E. Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday* **21**, 11 (2016).
16. Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. The rise of social bots. *Commun. ACM* **59**, 96–104 (2016).
17. Lazer, D. The rise of the social algorithm. *Science* **348**, 1090–1091 (2015).
18. Tufekci, Z. Engineering the public: big data, surveillance and computational politics. *First Monday* **19**, 7 (2014).
19. Lee, T.-S. & Chen, I.-F. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Syst. Appl.* **28**, 743–752 (2005).
20. Roszbach, K. Bank lending policy, credit scoring, and the survival of loans. *Rev. Econ. Stat.* **86**, 946–958 (2004).
21. Huang, C.-L., Chen, M.-C. & Wang, C.-J. Credit scoring with a data mining approach based on support vector machines. *Expert Syst. Appl.* **33**, 847–856 (2007).
22. Tsai, C.-F. & Wu, J.-W. Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Syst. Appl.* **34**, 2639–2649 (2008).
23. Chen, L. & Wilson, C. Observing algorithmic marketplaces in-the-wild. *SIGecom Exch.* **15**, 34–39 (2017).
24. Chen, L., Mislove, A. & Wilson, C. An empirical analysis of algorithmic pricing on Amazon marketplace. In *Proc. 25th International Conference on World Wide Web* 1339–1349 (International World Wide Web Conferences Steering Committee, 2016).
25. Hannák, A. et al. Bias in Online freelance marketplaces: evidence from TaskRabbit and Fiverr. In *Proc. ACM Conference on Computer Supported Cooperative Work and Social Computing* 1914–1933 (2017).
26. Cartlidge, J., Szostek, C., De Luca, M. & Cliff, D. Too fast too furious—faster financial-market trading agents can give less efficient markets. In *Proc. 4th International Conference on Agents and Artificial Intelligence* 126–135 (2012).
27. Kearns, M., Kulesza, A. & Nevmyvaka, Y. Empirical limitations on high-frequency trading profitability. *J. Trading* **5**, 50–62 (2010).
28. Wellman, M. P. & Rajan, U. Ethical issues for autonomous trading agents. *Minds Mach.* **27**, 609–624 (2017).
29. Farmer, J. D. & Skouras, S. An ecological perspective on the future of computer trading. *Quant. Finance* **13**, 325–346 (2013).
30. Perry, W. L., McInnis, B., Price, C. C., Smith, S. & Hollywood, J. S. *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations* (RAND, 2013).
31. Bonnefon, J.-F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576 (2016).
32. Kooti, F. et al. Analyzing Uber's ride-sharing economy. In *Proc. 26th International Conference on World Wide Web* 574–582 (International World Wide Web Conferences Steering Committee, 2017).
33. Zeng, X., Fapojuwo, A. O. & Davies, R. J. Design and performance evaluation of voice activated wireless home devices. *IEEE Trans. Consum. Electron.* **52**, 983–989 (2006).
34. Hendriks, B., Meerbeek, B., Boess, S., Pauws, S. & Sonneveld, M. Robot vacuum cleaner personality and behavior. *Int. J. Soc. Robot.* **3**, 187–195 (2011).
35. Hitsch, G. J., Hortaçsu, A. & Ariely, D. Matching and sorting in online dating. *Am. Econ. Rev.* **100**, 130–163 (2010).
36. Finkel, E. J., Eastwick, P. W., Karney, B. R., Reis, H. T. & Sprecher, S. Online dating: a critical analysis from the perspective of psychological science. *Psychol. Sci. Public Interest* **13**, 3–66 (2012).
37. Park, H. W., Rosenberg-Kima, R., Rosenberg, M., Gordon, G. & Breazeal, C. Growing growth mindset with a social robot peer. In *Proc. 2017 ACM/IEEE International Conference on Human–Robot Interaction* 137–145 (ACM, 2017).
38. Bemelmans, R., Gelderblom, G. J., Jonker, P. & de Witte, L. Socially assistive robots in elderly care: a systematic review into effects and effectiveness. *J. Am. Med. Dir. Assoc.* **13**, 114–120 (2012).
39. Shirado, H. & Christakis, N. A. Locally noisy autonomous agents improve global human coordination in network experiments. *Nature* **545**, 370–374 (2017). **In this human–machine hybrid study, the authors show that simple algorithms injected into human gameplay can improve coordination outcomes among humans**.
40. Pichai, S. AI at Google: Our Principles. *Google Blog* https://blog.google/topics/ai/ai-principles/ (2018).
41. Roff, H. M. The strategic robot problem: lethal autonomous weapons in war. *J. Mil. Ethics* **13**, 211–227 (2014).
42. Krishnan, A. *Killer Robots: Legality and Ethicality of Autonomous Weapons* (Routledge, 2016).
43. Voosen, P. The AI detectives. *Science* **357**, 22–27 (2017).
44. Szegedy, C. et al. Intriguing properties of neural networks. Preprint at https://arxiv.org/abs/1312.6199 (2013).
45. Zhang, Q.-S. & Zhu, S.-C. Visual interpretability for deep learning: a survey. *Front. Inf. Technol. Electronic Eng.* **19**, 27–39 (2018).
46. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. Preprint at https://arxiv.org/abs/1702.08608 (2017).
47. Gebru, T. et al. Datasheets for datasets. Preprint at https://arxiv.org/abs/1803.09010 (2018).
48. Mitchell, M. et al. Model cards for model reporting. Preprint at https://arxiv.org/abs/1810.03993 (2018).
49. Lakkaraju, H., Kamar, E., Caruana, R. & Horvitz, E. Identifying unknown unknowns in the open world: representations and policies for guided exploration. In *Proc. 31st Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence* 2 (2017).
50. Johnson, N. et al. Abrupt rise of new machine ecology beyond human response time. *Sci. Rep.* **3**, 2627 (2013).
51. Appel, K., Haken, W. & Koch, J. Every planar map is four colorable. Part II: reducibility. *Illinois J. Math.* **21**, 491–567 (1977).

52. Appel, K. & Haken, W. Every planar map is four colorable. Part I: discharging. *Illinois J. Math.* **21**, 429–490 (1977).
53. Westlund, J. M. K., Park, H. W., Williams, R. & Breazeal, C. Measuring young children's long-term relationships with social robots. In *Proc. 17th ACM Conference on Interaction Design and Children* 207–218 (ACM, 2018).
54. Lorenz, T., Weiss, A. & Hirche, S. Synchrony and reciprocity: key mechanisms for social companion robots in therapy and care. *Int. J. Soc. Robot.* **8**, 125–143 (2016).
55. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
This study examines the complex hybrid ecology of bots and humans on Twitter and finds that humans spread false information at higher rates than bots.
56. Lazer, D. M. J. et al. The science of fake news. *Science* **359**, 1094–1096 (2018).
57. Roberts, M. E. *Censored: Distraction and Diversion Inside China's Great Firewall* (Princeton Univ. Press, 2018).
58. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. Algorithmic decision making and the cost of fairness. In *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 797–806 (ACM, 2017).
59. Kleinberg, J., Mullainathan, S. & Raghavan, M. Inherent trade-offs in the fair determination of risk scores. Preprint at https://arxiv.org/abs/1609.05807 (2016).
60. Buolamwini, J. & Gebru, T. Gender shades: intersectional accuracy disparities in commercial gender classification. In *Proc. 1st Conference on Fairness, Accountability and Transparency* (eds Friedler, S. A. & Wilson, C.) **81**, 77–91 (PMLR, 2018).
61. Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. & Kalai, A. T. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proc. Advances in Neural Information Processing Systems* 4349–4357 (2016).
62. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
63. Sweeney, L. Discrimination in online ad delivery. *Queueing Syst.* **11**, 10 (2013).
64. Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C. & Venkatasubramanian, S. Runaway feedback loops in predictive policing. Preprint at https://arxiv.org/abs/1706.09847 (2017).
65. Angwin, J., Larson, J., Mattu, S. & Kirchner, L. Machine bias. *ProPublica* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016).
66. Chouldechova, A., Benavides-Prado, D., Fialko, O. & Vaithianathan, R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proc. 1st Conference on Fairness, Accountability and Transparency* (eds Friedler, S. A. & Wilson, C.) **81**, 134–148 (PMLR, 2018).
67. Jennings, N. R. et al. Human–agent collectives. *Commun. ACM* **57**, 80–88 (2014).
68. Campbell, M., Hoane, A. J. & Hsu, F.-H. Deep blue. *Artif. Intell.* **134**, 57–83 (2002).
69. Schaeffer, J. et al. Checkers is solved. *Science* **317**, 1518–1522 (2007).
70. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
71. Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
72. Bowling, M., Burch, N., Johanson, M. & Tammelin, O. Heads-up limit hold'em poker is solved. *Science* **347**, 145–149 (2015).
73. Bellemare, M. G., Naddaf, Y., Veness, J. & Bowling, M. The arcade learning environment: an evaluation platform for general agents. *J. Artif. Intell. Res.* **47**, 253–279 (2013).
74. Wellman, M. P. et al. Designing the market game for a trading agent competition. *IEEE Internet Comput.* **5**, 43–51 (2001).
75. Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I. & Osawa, E. RoboCup: the robot world cup initiative. In *Proc. 1st International Conference on Autonomous Agents* 340–347 (ACM, 1997).
76. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
77. Lin, T.-Y. et al. Microsoft COCO: common objects in context. In *Proc. European Conference on Computer Vision* (eds Fleet, D. et al.) 8693, 740–755 (Springer International Publishing, 2014).
78. Davis, J. & Goadrich, M. The relationship between precision–recall and ROC curves. In *Proc. 23rd International Conference on Machine Learning* 233–240 (ACM, 2006).
79. van de Sande, K. E. A., Gevers, T. & Snoek, C. G. M. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 1582–1596 (2010).
80. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting on Association for Computational Linguistics* 311–318 (Association for Computational Linguistics, 2002).
81. Zhou, Z., Zhang, W. & Wang, J. Inception score, label smoothing, gradient vanishing and -log(D(x)) alternative. Preprint at https://arxiv.org/abs/1708.01729 (2017).
82. Epstein, Z. et al. Closing the AI knowledge gap. Preprint at https://arxiv.org/abs/1803.07233 (2018).
83. Tinbergen, N. On aims and methods of ethology. *Ethology* **20**, 410–433 (1963).
84. Nesse, R. M. Tinbergen's four questions, organized: a response to Bateson and Laland. *Trends Ecol. Evol.* **28**, 681–682 (2013).
85. Das, R., Hanson, J. E., Kephart, J. O. & Tesauro, G. Agent–human interactions in the continuous double auction. In *Proc. 17th International Joint Conference on Artificial Intelligence* 1169–1178 (Lawrence Erlbaum, 2001).
86. Deng, Y., Bao, F., Kong, Y., Ren, Z. & Dai, Q. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Trans. Neural Netw. Learn. Syst.* **28**, 653–664 (2017).
87. Galceran, E., Cunningham, A. G., Eustice, R. M. & Olson, E. Multipolicy decision-making for autonomous driving via changepoint-based behavior prediction: theory and experiment. *Auton. Robots* **41**, 1367–1382 (2017).
88. Ribeiro, M. T., Singh, S. & Guestrin, C. Why should I trust you? Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (ACM, 2016).
89. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. SmoothGrad: removing noise by adding noise. Preprint at https://arxiv.org/abs/1706.03825 (2017).
90. Nevmyvaka, Y., Feng, Y. & Kearns, M. reinforcement learning for optimized trade execution. In *Proc. 23rd International Conference on Machine Learning* 673–680 (ACM, 2006).
91. Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L. & Konstan, J. A. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proc. 23rd International Conference on World Wide Web* 677–686 (ACM, 2014).
92. Dalvi, N. & Domingos, P. Mausam, Sanghai, S. & Verma, D. Adversarial classification. In *Proc. Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 99–108 (ACM, 2004).
93. Globerson, A. & Roweis, S. Nightmare at test time: robust learning by feature deletion. In *Proc. 23rd International Conference on Machine Learning* 353–360 (ACM, 2006).
94. Biggio, B. et al. Evasion attacks against machine learning at test time. In *Proc. Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 387–402 (Springer, 2013).
95. Tramèr, F. et al. Ensemble adversarial training: attacks and defenses. Preprint at https://arxiv.org/abs/1705.07204 (2017).
96. Parkes, D. C. & Wellman, M. P. Economic reasoning and artificial intelligence. *Science* **349**, 267–272 (2015).
97. Wagner, A. *Robustness and Evolvability in Living Systems* (Princeton Univ. Press, 2013).
98. Edwards, H. & Storkey, A. Censoring representations with an adversary. Preprint at https://arxiv.org/abs/1511.05897 (2015).
99. Zemel, R., Wu, Y., Swersky, K., Pitassi, T. & Dwork, C. learning fair representations. *In Proc. International Conference on Machine Learning* 325–333 (2013).
100. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C. & Venkatasubramanian, S. Certifying and removing disparate impact. In *Proc. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 259–268 (ACM, 2015).
101. Cully, A., Clune, J., Tarapore, D. & Mouret, J.-B. Robots that can adapt like animals. *Nature* **521**, 503–507 (2015).
This study characterizes a robot driven by an adaptive algorithm that mimics the adaptation and behaviours of animals.
102. Bongard, J., Zykov, V. & Lipson, H. Resilient machines through continuous self-modeling. *Science* **314**, 1118–1121 (2006).
103. Leibo, J. Z. et al. Psychlab: a psychology laboratory for deep reinforcement learning agents. Preprint at https://arxiv.org/abs/1801.08116 (2018).
In this study, the authors use behavioural tools from the life sciences in the study of machine behaviours.
104. Subrahmanian, V. S. et al. The DARPA Twitter bot challenge. Preprint at https://arxiv.org/abs/1601.05140 (2016).
105. Carrascosa, J. M., Mikians, J., Cuevas, R., Erramilli, V. & Laoutaris, N. I. Always feel like somebody's watching me: measuring online behavioural advertising. In *Proc. 11th ACM Conference on Emerging Networking Experiments and Technologies* **13** (ACM, 2015).
106. Datta, A., Tschantz, M. C. & Datta, A. Automated Experiments on Ad Privacy Settings. *Proc. Privacy Enhancing Technologies* **2015**, 92–112 (2015).
107. Giusti, A. et al. A machine learning approach to visual perception of forest trails for mobile robots. *IEEE Robot. Autom. Lett.* **1**, 661–667 (2016).
108. Berdahl, A., Torney, C. J., Ioannou, C. C., Faria, J. J. & Couzin, I. D. Emergent sensing of complex environments by mobile animal groups. *Science* **339**, 574–576 (2013).
109. Couzin, I. D. et al. Uninformed individuals promote democratic consensus in animal groups. *Science* **334**, 1578–1580 (2011).
110. Rubenstein, M., Cornejo, A. & Nagpal, R. Programmable self-assembly in a thousand-robot swarm. *Science* **345**, 795–799 (2014).
111. Kernbach, S., Thenius, R., Kernbach, O. & Schmickl, T. Re-embodiment of honeybee aggregation behavior in an artificial micro-robotic system. *Adapt. Behav.* **17**, 237–259 (2009).
112. Bak, P., Chen, K. & Creutz, M. Self-organized criticality in the 'Game of Life'. *Nature* **342**, 780–782 (1989).
113. Tsvetkova, M., García-Gavilanes, R., Floridi, L. & Yasseri, T. Even good bots fight: the case of Wikipedia. *PLoS ONE* **12**, e0171774 (2017).
114. Lazaridou, A., Peysakhovich, A. & Baroni, M. Multi-agent cooperation and the emergence of (natural) language. Preprint at https://arxiv.org/abs/1612.07182 (2016).
115. Budish, E., Cramton, P. & Shim, J. The high-frequency trading arms race: frequent batch auctions as a market design response. *Q. J. Econ.* **130**, 1547–1621 (2015).
116. Kirilenko, A. A. & Lo, A. W. Moore's law versus Murphy's law: algorithmic trading and its discontents. *J. Econ. Perspect.* **27**, 51–72 (2013).

117. Menkveld, A. J. The economics of high-frequency trading: taking stock. *Annu. Rev. Financ. Econ.* **8**, 1–24 (2016).
118. Mønsted, B., Sapieżyński, P., Ferrara, E. & Lehmann, S. Evidence of complex contagion of information in social media: an experiment using Twitter bots. *PLoS ONE* **12**, e0184148 (2017).
This study presents an experimental intervention on Twitter using bots and provides evidence that information diffusion is most accurately described by complex contagion.
119. Bainbridge, L. Ironies of automation. *Automatica* **19**, 775–779 (1983).
120. Jeong, S., Breazeal, C., Logan, D. & Weinstock, P. Huggable: the impact of embodiment on promoting socio-emotional interactions for young pediatric inpatients. In *Proc. 2018 CHI Conference on Human Factors in Computing Systems* 495 (ACM, 2018).
121. Kory Westlund, J. M. et al. Flat vs. expressive storytelling: young children's learning and retention of a social robot's narrative. *Front. Hum. Neurosci.* **11**, 295 (2017).
122. Salisbury, E., Kamar, E. & Morris, M. R. Toward scalable social alt text: conversational crowdsourcing as a tool for refining vision-to-language technology for the blind. *Proc. 5th AAAI Conference on Human Computation and Crowdsourcing* (2017).
123. Awad, E. et al. The Moral Machine experiment. *Nature* **563**, 59–64 (2018).
124. Dietvorst, B. J., Simmons, J. P. & Massey, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* **144**, 114–126 (2015).
125. Gray, K. & Wegner, D. M. Feeling robots and human zombies: mind perception and the uncanny valley. *Cognition* **125**, 125–130 (2012).
126. Brynjolfsson, E. & Mitchell, T. What can machine learning do? Workforce implications. *Science* **358**, 1530–1534 (2017).
127. Christiano, P. F. et al. Deep reinforcement learning from human preferences. In *Proc. Advances in Neural Information Processing Systems* 30 (eds Guyon, I. et al.) 4299–4307 (Curran Associates, 2017).
128. Tsvetkova, M. et al. Understanding human–machine networks: a cross-disciplinary survey. *ACM Comput. Surv.* **50**, 12:1–12:35 (2017).
129. Hilbert, M., Ahmed, S., Cho, J., Liu, B. & Luu, J. Communicating with algorithms: a transfer entropy analysis of emotions-based escapes from online echo chambers. *Commun. Methods Meas.* **12**, 260–275 (2018).
130. Kramer, A. D. I., Guillory, J. E. & Hancock, J. T. Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl Acad. Sci. USA* **111**, 8788–8790 (2014).
131. Kamar, E., Hacker, S. & Horvitz, E. Combining human and machine intelligence in large-scale crowdsourcing. in *Proc. 11th International Conference on Autonomous Agents and Multiagent Systems* 467–474 (International Foundation for Autonomous Agents and Multiagent Systems, 2012).
132. Jackson, M. *The Human Network: How Your Social Position Determines Your Power, Beliefs, and Behaviors* (Knopf Doubleday, 2019).
133. Crandall, J. W. et al. Cooperating with machines. *Nat. Commun.* **9**, 233 (2018).
This study examines algorithmic cooperation with humans and provides an example of methods that can be used to study the behaviour of human–machine hybrid systems.
134. Wang, D., Khosla, A., Gargeya, R., Irshad, H. & Beck, A. H. Deep learning for identifying metastatic breast cancer. Preprint at https://arxiv.org/abs/1606.05718 (2016).
135. Pentland, A. *Social Physics: How Social Networks Can Make Us Smarter* (Penguin, 2015).
136. Lazer, D. et al. Computational social science. *Science* **323**, 721–723 (2009).
137. Aharony, N., Pan, W., Ip, C., Khayal, I. & Pentland, A. Social fMRI: investigating and shaping social mechanisms in the real world. *Pervasive Mobile Comput.* **7**, 643–659 (2011).
138. Ledford, H. How to solve the world's biggest problems. *Nature* **525**, 308–311 (2015).
139. Bromham, L., Dinnage, R. & Hua, X. Interdisciplinary research has consistently lower funding success. *Nature* **534**, 684–687 (2016).
140. Kleinberg, J. & Oren, S. Mechanisms for (mis)allocating scientific credit. In *Proc. 43rd Annual ACM Symposium on Theory of Computing* 529–538 (ACM, 2011).
141. Kannel, W. B. & McGee, D. L. Diabetes and cardiovascular disease. The Framingham study. *J. Am. Med. Assoc.* **241**, 2035–2038 (1979).
142. Krafft, P. M., Macy, M. & Pentland, A. Bots as virtual confederates: design and ethics. In *Proc. 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* 183–190 (ACM, 2017).
143. Meyer, M. N. Two cheers for corporate experimentation: The A/B illusion and the virtues of data-driven innovation. *Colorado Technol. Law J.* **13**, 273 (2015).
144. Xing, X. et al. Take this personally: pollution attacks on personalized services. In *Proc. 22nd USENIX Security Symposium* 671–686 (2013).
145. Patel, K. Testing the limits of the First Amendment: how a CFAA prohibition on online antidiscrimination testing infringes on protected speech activity. *Columbia Law Rev.* https://doi.org/10.2139/ssrn.3046847 (2017).

Acknowledgements I.R. received funding from the Ethics & Governance of Artificial Intelligence Fund; J.B. received funding from NSF awards INSPIRE-1344227 and BIGDATA-1447634, DARPA's Lifelong Learning Machines program, ARO contract W911NF-16-1-0304; J.-F.B. from the ANR-Labex IAST; N.A.C. a Pioneer Grant from the Robert Wood Johnson Foundation; I.D.C. received funding from the NSF (IOS-1355061), the ONR (N00014-09-1-1074 and N00014-14-1-0635), the ARO (W911NG-11-1-0385 and W911NF14-1-0431), the Struktur- und Innovationsfunds für die Forschung of the State of Baden-Württemberg, the Max Planck Society and the DFG Centre of Excellence 2117 "Centre for the Advanced Study of Collective Behaviour" (422037984); D.L. received funding from the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) under contract 2017-17061500006; J.B.T. received funding from the Center for Brains, Minds and Machines (CBMM) under NSF STC award CCF–1231216; M.W. received funding from the Future of Life Institute. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA or the US Government.

Author contributions I.R., M.C. and N.O. conceived the idea, produced the figures and drafted the manuscript. All authors contributed content, and refined and edited the manuscript.

Competing interests The authors declare no competing interests.