Cristianini, N., Scantamburlo, T., & Ladyman, J. A. C. (2021). The social turn of artificial intelligence. *AI and Society*, 0-0. [0]. https://doi.org/10.1007/s00146-021-01289-8

Publisher's PDF, also known as Version of record

License (if available):
CC BY

Link to published version (if available):
10.1007/s00146-021-01289-8

Link to publication record in Explore Bristol Research
PDF-document

This is the final published version of the article (version of record). It first appeared online via Springer at https://doi.org/10.1007/s00146-021-01289-8 .Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/

AI & SOCIETY
https://doi.org/10.1007/s00146-021-01289-8

ORIGINAL ARTICLE

The social turn of artificial intelligence

Nello Cristianini[1] · Teresa Scantamburlo[2] · James Ladyman[3]

Received: 23 June 2021 / Accepted: 13 September 2021
© The Author(s) 2021

Abstract

Social machines are systems formed by material and human elements interacting in a structured way. The use of digital platforms as mediators allows large numbers of humans to participate in such machines, which have interconnected AI and human components operating as a single system capable of highly sophisticated behaviour. Under certain conditions, such systems can be understood as autonomous goal-driven agents. Many popular online platforms can be regarded as instances of this class of agent. We argue that autonomous social machines provide a new paradigm for the design of intelligent systems, marking a new phase in AI. After describing the characteristics of goal-driven social machines, we discuss the consequences of their adoption, for the practice of artificial intelligence as well as for its regulation.

Keywords Intelligent agents · Social machines · Artificial intelligence · Human–computer interaction · Cybernetics · Autonomous agents · Teleology

1 Introduction

Over the past decade, digital platforms have become social intermediaries, often powered by recommendation algorithms, and incorporated into online shops, news aggregators, marketplaces, service sharing, video streaming and social networks. The automation of high-quality decisions about complex situations has enabled the replacement of human operators and displaced many forms of human culture. Some digital platforms have over a billion users (e.g., Facebook reported 1.9 billion daily average users, and 2.9 billion monthly average users, in June 2021; Facebook 2021).

To discuss such systems, we use the concept of "social machines", which originates with Berners-Lee and Fischetti 1999, although we change somewhat its original emphasis. A machine is an apparatus composed of specialised material parts interacting together to do a particular type of work. There may be very different substrates of those parts, which can include—for example—hydraulic, electrical, and mechanical components. A social machine (SM) is a machine in which some components, carrying out specific subtasks, are human beings (whom we call 'participants'). In the case of present interest, the interaction of human participants is mediated by a digital infrastructure that provides them with information and constrains the actions and communications that they can take. There has been considerable work in this area over the past few years (Shadbolt et al. 2019; Smart and Shadbolt 2014). However, we somewhat reframe the notions of autonomy and teleology used in that literature, making them more apt and specific.

This article identifies a subclass of SMs that can be regarded as autonomous and goal-driven, and therefore considered as intelligent agents. We argue that some of the digital platforms mentioned above can be understood in this way, and we do so to examine some of their consequences for society.

We assume a restricted definition of intelligence as intelligent behaviour, as commonly assumed in modern artificial intelligence (AI) and biology (Turing, 1948; Russell and

✉ Nello Cristianini
nello.cristianini@bristol.ac.uk

Teresa Scantamburlo
teresa.scantamburlo@unive.it

James Ladyman
James.Ladyman@bristol.ac.uk

[1] Departments of Engineering Mathematics and Computer Science, University of Bristol, Ada Lovelace Building, University Walk, Bristol BS8 1TW, UK

[2] European Centre for Living Technology, Università Cà Foscari Venezia, Venezia, Italy

[3] Department of Philosophy, University of Bristol, Bristol, UK

Published online: 04 October 2021

Norvig 2002; McFarland and Bösser 2002). We focus on goal-driven intelligent behaviour, assuming that artificial, biological, and even social systems can behave intelligently when they make choices that are better than random at achieving goals. This behaviour is also sometimes called `purposeful' or `teleological' in cybernetics (Rosenblueth et al. 1943; Wiener 1948), and `rational' in economics and game theory (von Neumann and Morgenstern 1953). On this view, ant colonies pursuing their goals of building nests display intelligent behaviour, as do online recommender systems pursuing their goals of user engagement.

We argue that AI-enabled social machines endowed with autonomous and teleological behaviour exist, and that there is no need for their goals to be aligned with those of their participants. After defining some key terms and giving some examples, we discuss some implications of this observation for both Artificial Intelligence and for Society.

Typical examples of such machines include systems based on collaborative filtering, commonly used in the automatic recommendation of media content online, where individual participants have the aim of finding the information they need, while the overall system is pursuing a different goal, such as maximising advertising revenue. Other examples include online marketplaces, where users can bid for items, or cooperative video games. These systems are mediated by statistical or learning algorithms and informed by interactions and iterated feedback among large numbers of users.

These new macroscopic agents are not just informed by observing the behaviour of their users, they co-opt them as participants to delegate to them delicate decision-making tasks. A steady stream of elementary choices is elicited, gathered, and repackaged to inform the behaviour of the system, enabling it to make decisions for which there is currently no algorithm.

To keep the discussion simple, we adopt here the definition of artificial intelligence used by the European Commission (EU HLEG AI 2019): "Artificial intelligence (AI) refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to predefined parameters) to achieve the given goal."

In other words, we define intelligent behaviour as "deciding the best actions to achieve a given goal, either in the physical or the digital world". We, therefore, rephrase the question posed by Alan Turing in 1948: "whether it is possible for machinery to show intelligent behaviour" (Turing 1948, p. 1) as "whether it is possible for SMs to show autonomous and purposeful behaviour". Sections 2, 3 and 4 are devoted to answering this question.

We observe that under these definitions, a collaborative filtering algorithm together with its users, such as the one recommending videos on YouTube, qualifies as an autonomous and goal-driven SM, as well as an intelligent agent. Its behaviour is not determined solely by an algorithm, but by its application to the usage patterns of myriad users, none of which is in the position to control it. There is no homunculus in charge of such systems, which can be likened to ant colonies or the human brain, because adaptive behaviour emerges from the interactions of their parts involving many iterations of feedback between them. The behaviour of such SMs results from the interaction of components, some of which happen to be human users (who do not need to be willing or aware participants). The seat of intelligence in those systems is neither in the algorithm nor in the participants but in the interaction of all the components (See for example Dennett and Hofstaedter 1981; Boden 1990; Kurzweil 2002; Ladyman and Wiesner 2020 discuss the emergent intelligence of social insects and other aspects of complex systems relevant to social machines including the importance of iterated interaction and feedback).

The consequences of these observations for the practice of intelligent systems design, as well as for their legal regulation, are important and considered in Sects. 5 and 6.

2 Teleological (goal-driven) agents

An **agent** is any system that can choose among multiple actions on the environment, in the sense that there is at least one variable in the environment the value of which depends on its choice. In general, there are various possible incompatible actions among which a choice is made, and the state of the environment that results probabilistically depends on the action performed. This definition of agent is general and includes, for example, organisms, artifacts, and organizations.

An agent is **autonomous** if it chooses its actions "by itself", in the sense that its internal processes and states determine the choices in a way that depends on its current and previous interactions with the environment. An agent which is not autonomous, but whose actions are chosen by an external controller (or process), is **heteronomous**. Examples of autonomous agents include software agents whose actions are chosen by them in response to the specific situation, as in the familiar case of a personalised video recommendation portal; an example of a heteronomous agent may be a bureaucracy worker always performing the same actions as directed and without regard for the overall outcome.

A **teleological** (or goal-driven, or purposeful) agent is an autonomous agent whose behaviour can be modelled as choosing based on the expected utility of the outcomes of the different actions it can perform. These actions do not necessarily need to be optimal, nor to have deterministic effects on the environment, but on average the agent's behaviour must do better than random in achieving some outcome, which can then be thought of as the agent's goal and hence as having utility for it. Of course, the same behaviour may be teleological in one environment, and not teleological in another environment (and teleological behaviour is not possible if the effects of actions on the environment are completely random). This class of agents that can fruitfully be described in non-causal language are called purposeful, as used in the classic paper (Rosenblueth et al. 1943): "The term purposeful is meant to denote that the act or behaviour may be interpreted as directed to the attainment of a goal."

Typically, agents must have information about the state of their environment to calculate the expected utility of different actions, in other words they need to be able to sense. An agent "senses" the environment if at least one variable of the agent's internal state can be coupled with at least one variable in the state of the environment and then used in information-processing about what choice to make.

The fact that the overall goal is normally set by the external designer of the agent does not necessarily reduce the autonomy of the agent, in that the agent still chooses its own actions. For example, all American eels migrate to the Sargasso Sea each year, but individual eels choose the necessary actions based on their specific situation, making them autonomous agents.

For the rest of this paper, we refer to any agents that autonomously make informed decisions in pursuit of goals as "intelligent", and we use interchangeably the expressions "goal-driven", "purposeful", and "teleological" to describe behaviour. Note that pursuing a goal might include maximising utility or maintaining homeostasis.

3 Social machines

SMs are particular types of machine in which some components, carrying out specific subtasks, are human beings (whom we call 'participants'). While the authors of (Berners-Lee and Fischetti 1999) presuppose a digital infrastructure to mediate the activities of the participants, the mediator could be a physical bureaucracy, or machinery. What is key is that the interaction among the participants is mediated by the infrastructure via a standardised interface that can also constrain the options that are available to the participants and how they communicate.

Mechanisms incorporating 'participants' can be found in multiple areas and include assembly lines, bureaucracies, auctions, markets, voting schemes, product delivery services, games, peer production, crowdsourcing, and so on. Their interaction can be mediated by forms, ballots, purchase orders, and IT systems. In this study, we are particularly interested in cases where an intelligent software infrastructure mediates the interactions among the various participants, constraining and standardising their actions, monitoring performance, and possibly also assigning incentives.

The most recent generation of SMs builds upon web-based infrastructures, which affords them many possibilities that were not available to previous versions such as bureaucracies or assembly lines. Not only does this allow them to include billions of participants, creating completely new dynamics, but also to rapidly transfer information from one part of the system to another.

Two different examples show the range of possibilities and illustrate where our emphasis differs from the perspective normally taken in the study of SMs (Shadbolt et al. 2019).

Online crowdsourcing services, such as Amazon's Mechanical Turk, operate as modern assembly lines, where thousands of workers are active at any given time (Difallah et al. 2018) to perform well specified tasks that might be difficult to automate. In this type of modern SM participants can be asked—for example—to perform very specific tasks, such as tagging photos and videos, entering data from handwritten receipts, categorizing images, answering questions and so forth. The tasks are tightly specified from outside the SM, which is mostly used to distribute the tasks to many participants and monitor their performance. They may not be aware of the overall purposes of the "customer", but they do understand what is the task that they are completing. The system is executing externally set tasks.

Such cases stand in sharp contrast with recommender systems (Ricci et al. 2011), such as those used in video streaming and social media, and which are typically mediated by collaborative filtering algorithms. In this case, as a user accesses some type of content, their spontaneous activity is monitored, so that the system can infer detailed information about both its users and its catalogue of items. The system is using this information both to steer the individual user towards clicking on certain content, and to learn how to treat new users or new content. Maintaining high performance is important for the constant co-optation of participants, so that the system approaches the condition of an autopoietic process (Maturana and Varela 1991).

The information is not explicitly given to the system by the users, rather it comes through "implicit feedback" (Oard and Kim 1998). In the example of YouTube, for each choice made by a user, a wealth of information is recorded: such as whether a user is logged or not, IDs of watched videos, watch time, clicked and unclicked video impressions, time since last watched video, and so on (Covington 2016). With each action, users disclose personal preferences and interests to the system (Burr et al. 2018) and without their contribution the whole machine would not be able to do its job, which is to offer personalised recommendations that lead to some kind of user engagement (for example, the user watching adverts).

The system is pursuing an externally set goal (perhaps to increase traffic and advertising revenue) but is not executing external given instructions: it is making its own choices, while pursuing its given goal.

When considering the relation between participants and the machine of which they are part we refer to the behaviour of participants and that of the system, as the `micro' and 'macro' levels, respectively. The macro-level behaviour of the SM depends on the micro-level behaviour of each participant, which itself depends on the behaviour of the non-human and other human components of the SM. Agents at both levels can be autonomous or heteronomous.

Crowdsourcing SMs are often heteronomous, in that someone external to the system decides which specific tasks need to be completed, while recommender systems are autonomous. It is this second class of systems with which we are concerned here.

4 Teleological social machines

In the light of the previous section, it is clear that SMs can be autonomous and teleological, and can therefore meet the conditions for intelligent behaviour in Sect. 2. A simple example shows how an autonomous SM can be created whose effect is to pursue a specific goal (at the macro-level), which its participants (at the micro-level) cannot control.

4.1 Examples

The ESP game consists of an online platform where thousands of pairs of players are randomly matched to play a sort of guessing game without communicating with one another. The objective is to guess which label the other partner will assign to a given image that both players can see. Since coordination among players is ruled out by the design of the interface and the randomised matching, for each player the optimal strategy to maximise their own score is to guess the most probable word describing each image, and a spontaneous consequence of multiple participants playing this game is the production of high-quality annotation for large dataset. Experiments with the ESP game showed that in 1 month 5,000 people can produce accurate labels for more than 400,000,000 images (von Ahn and Dabbish 2004; von Ahn et al. 2002). While the goal of individual players is playing a video game, the overall effect at the macro-level is to generate increasingly reliable data annotation. This spontaneous tendency towards a state is goal-driven behaviour in the sense of Sect. 2, and it emerges from the structure of the game.

The example of the ESP game makes a number of points clear. Most importantly, the goals of participants may be different from the goals of the system, participants do not even need to be aware that they are part of a system which has goals. The spontaneous drift of the system towards a state where the data is increasingly well annotated is an instance of what Adam Smith called "the invisible hand" that guides the markets while the participants pursue their own goals. The case of this game has other features in common with markets: it is an example of a "strategic game", where the benefits of actions depend on the actions taken by other players. These can be either competitive (as in markets) or cooperative (as in the ESP game). In both cases, the "invisible hand" can guide the system towards a macro-level goal.

This class of SMs is more specific than those commonly studied, typically based on the principle of "crowdsourcing" in which a task is broken into smaller tasks, and then distributed among willing participants, much like for the Mechanical Turk described above (Shadbolt et al. 2019; Smart et al. 2014; Berners-Lee and Fischetti 1999).

An example of an autonomous SM which is not based on a strategic game is the recommender system, perhaps based on collaborative filtering, which is used to suggest books, videos, and even romantic partners. These systems work by leveraging the micro-decisions of a large number of participants, and using them to steer towards a goal, but they are not requested to make strategic decisions, just to keep on choosing one item out of many proposed. It is this stream of choices that generates the necessary information for the macro-level system to steer autonomously towards its goal (for example, that of retaining participants, and maximising the time they spend on it) (Covington 2016).

4.2 The purpose of a system

The notion of rational agents naturally connects to the idea of a purpose in systems-cybernetics. The cyberneticist Stafford Beer (Beer 2002) noticed how it is the behaviour of a system that reveals its "purposes", by definition. This observation was turned into a popular slogan, known as the POSIWID principle from the initials of each word: "the purpose of a system is what it does" (Beer 2002). In other words, if the emergent behaviour of a system has the net effect of pursuing a certain goal, then this is the system's purpose. We use this perspective, when we define the "telos" of an autonomous agent as the net effect of its choices. Using this language, we can say for example, that the interaction of a social machine such as eBay is framed in a way to identify the user willing to pay the highest price, when the goal of each user would be not to reveal that information, and to pay the lowest cost. The invisible hand of Adam Smith guides the system towards its telos, while individual users pursue their own.

4.3 Value alignment

Modelling both participants and the whole SM as rational agents directly introduces the question of the alignment between utility at micro-level and at macro-level. The lack of alignment between the goals of a user with those of an AI system has been identified as a problem in (HHadfield-Menell et al. 2017). For instance, YouTube and Facebook might have the goal of increasing the time that users spend with them, Amazon that of increasing the money they spend, while the users might have a different goal altogether.

4.4 Design principles for teleological social machines

In many circumstances, it would be desirable to program the behaviour of a SM at the macro-level to achieve a desired goal. For example, social network executives may want to pursue the goal of preventing outrageous contents. Unfortunately, no general method is known to do this, and only very specific subcases have been solved. The problem of translating a desired macro-behaviour into a system of incentives for participants is a largely open technical question, in some communities called the micro–macro-problem, particularly in the case of sociological theory (Alexander et al. 1987). In one specific subcase, however, this has been studied: for strategic games such as auctions, or even the ESP game, there is a part of Game Theory whose purpose is to design the rules of the game, so that the system pursues a chosen goal, while the participants can act "selfishly" (Börgers 2015). This subfield is called "mechanism design" and its techniques are used—for example—to design eBay auctions, which are an example of the (few) cases in which we can "program" Teleological SMs.

5 Discussion

The use of social machines to implement goal-driven agents marks a fundamental change from previous attempts at creating autonomous systems, one that bypasses many of the technical problems encountered before, but also creates a new set of challenges. On the upside, it allows the system to perform tasks such as establishing the market price of a good, finding the best content to engage a human user, or annotating images, without needing a deep understanding of the underlying subject matter. Crucial cognitive tasks can be decomposed into elementary problems and automatically delivered to human participants who solve them, thereby contributing to the solution of a larger problem that they do not need to see or understand.

On the downside, we have built intelligent agents that include human participants, and mediate a range of their activities, and the dynamics of such distributed systems can produce emergent phenomena such as echo chambers, price inflation, reputation bubbles and so on. These are difficult if not impossible to model in advance. If no humans are in control of these behaviours, not even the designers and certainly not the participants, who holds responsibility for them? For example, is it reasonable to expect that a company should avoid the circulation of fake news on a system that they cannot steer? Is it reasonable to allow an autonomous agent to operate in this way? These are crucial issues for the design and regulation of AI.

5.1 The social turn in the field of AI

The current recipe for the generation of autonomous goal-driven behaviour includes the use of human participants in larger systems which can only function when both machinery and humans interact in a tightly regulated manner. Recommender systems, online marketplaces, even spam filters and spell checkers are implicitly powered by the behaviour of multiple participants. Many of them can be regarded as autonomous social machines, combining a powerful infrastructure, clever algorithms, and millions of human users. The judgment and knowledge of participants, carefully elicited by a digital infrastructure (sometimes during routine interaction), is a central ingredient in the decisions that the system will make in the future. Some of these participants are willing and paid task workers, others are unaware users, motivated by a diversity of goals (e.g., watching videos, deleting spam or buying products). This combination of databases, learning algorithms, and participants is a SM, and is in many cases autonomous, in the sense of not being controlled by any other system, including its designer. Recent concerns about the fairness of intelligent systems have been traced to the social origins of their training data, showing that their behaviour may be less predictable and less controllable than their makers might have hoped (e.g., Cristianini and Scantamburlo 2019; Sutton et al. 2018; Jonauskaite et al. 2021; Caliskan et al. 2017).

5.2 Consequences for individuals

The emergence of AI-enabled SMs has the potential to affect individuals in various ways (e.g., see Burr et al. 2018, for effects on human autonomy, and Cristianini 2019, for possible effects on fairness). The key point is that in this technology the interactions that give rise to the agent are all mediated by the software infrastructure, and human participants contribute by interacting with it through a constrained interface. This relation, where the interface is framed and controlled by the designer and its contents are adapted by the software to each participant, makes the machines work but raises concerns for individuals. We explore two orders of concerns: for human autonomy (e.g., Cambridge Analytica) and for employment (Amazon Warehouse, Uber riders).

5.2.1 The power of mediators

An important consequence is the emergence of a number of workers who are directly managed by a software layer, via an interface which could be a phone app. Participants—such as delivery drivers—are monitored and rewarded by an algorithm, and do not necessarily need to be aware of the overall goals they are part of pursuing. This can give rise to both ethical and management considerations about wellbeing, with workers being forced by a mechanism to compete for the lowest wages, or nudged into working more than they would like, but also exposing them to the risk of being one day replaced by automated systems. Since their entire contribution is mediated by the software interface, the rest of the system would not be affected if they were replaced by a mechanism. This can involve Uber drivers (Wakabayashi and Conger 2018), Amazon warehouse workers (Soper 2018), and various types of translators, writers, data curators. Many of these workers are already now working mostly to train their mechanic replacements. For those participants that contribute for free, instead, there are concerns involving the risk of exploitation (Rosenblatt 2018), particularly if there is the risk of behavioural addiction to certain services such as social media (Zendle and Bowden-Jones 2019).

5.2.2 On choice architectures and nudging

The interaction of the participant with the machine is framed by the machine itself, the interfaces determine the affordances of the participant, shaping the architecture within which any choice is made. For any individual participant working in these conditions, life is framed by the information and the options made available by it. In some ways, it behaves like a Skinner Box. Three important considerations follow from this observation about choice architectures: the system can design "mechanisms" (in the game-theoretic sense we defined above), by setting the micro-rules that the autonomous participant needs to follow; the system can exploit known psychological effects known collectively as "nudging", to steer the decisions of the participants; and the system can extract some (rudimentary) psychometric information from the participant. The first point has been discussed above, the second point is addressed in Burr et al. (2018), which discusses a variety of interactions between a human user and an intelligent software agent and examples may range from coercion, to nudging and persuasion. Examples of nudging include the way information is presented to a user at the moment they are asked to make a choice such as accepting a cookie, selecting a method of payment, watching one more video: different ways to present the same options (called "choice architectures") are known to lead to different decisions. The third point is discussed in (Burr and Cristianini 2019) where a series of examples are provided, in which the interaction between users and AI systems can disclose psychometric information about the users (Kosinski et al. 2013), an effect that was made widely known by the Cambridge Analytics affair. The main idea has analogies with the standard Item Response Theory of psychology: if you can carefully design a series of questions, you can observe samples of behaviour that will reveal latent information about the user.

5.3 Consequences for society

As we increasingly adopt SMs as part of our social infrastructure, we have come to rely on them as powerful mediators for many social functions, and as a new mass medium. While their influence at an individual level has been discussed above, we need to consider the possible effects on society as a whole. As the delivery of media content is mediated by a SM aimed at pursuing its own goals, what are the effects on public opinion or the economy? The same can be said for the recommendation of products, and effects on markets.

There are two key macro-level questions: how the machine can affect society, and how it can control its own participants. The second concern is rather fundamental: the internal (e.g., homeostatic) feedback loops that the machine uses to coordinate its own participants (e.g., incentive schemes), determine their overall behaviour. The degree of alignment between the goals of the machine and those of the designers might depend on the emergent dynamics connecting all participants via the infrastructure.

In some cases, SMs can be seen as mediating the communications between participants and coordinating their behaviour. This can result in various feedback dynamics, for example echo chambers that could promote "fake news". This can be seen as a sort of swarming behaviour, with popularity bubbles for videos, news, and products, whose control is not currently well understood.

6 Conclusions

There is a class of social machines that satisfy a standard definition of intelligence and can be understood as autonomous and goal-driven agents. Various online platforms that are currently used for a range of applications belong to this class: social networks (where contents are personally recommended to maximise engagement), online shops (where products are recommended) and marketplaces (where products are recommended and prices are established by a social mechanism), are all based on the same principle, of leveraging the decisions of millions of participants to inform the macroscopic behaviour of the system. Various outsourcing and sharing services follow similar principles (e.g., in car sharing, where human participants are matched with the customers by a similar mechanism). These systems do not just observe humans, they co-opt them, to be able to elicit the information needed for their operations. Human participants have no more control over the behaviour of the macro-system than do the cogs of a machine.

Understanding this class of systems is essential for their regulation, and this article proposes a unified terminology based on standard operational conceptions of intelligence and goal-driven behaviour. This conceptual framework can be useful for discussion of both design and of regulation. Thinking in terms of micro- and macro-levels, heteronomous and autonomous agents, participants and SMs, can help us make sense of this form of technology and the social challenges that it poses.

As early as 1960, Wiener suggested some moral problems "caused by the simultaneous action of the machine and the human being in a joint enterprise". These include the issues of individuals operating at a slower time scale and the creation of more opaque systems: "The result of a programming technique of automatization is to remove from the mind of the designer and operator an effective understanding of many of the stages by which the machine comes to its conclusions and of what the real tactical intentions of many of its operations may be" (Wiener 1960, p. 1358; see also (Wiener 1954)).

We have reached a point in the evolution of AI where the question of Alan Turing "whether machines are capable of intelligent behaviour" interacts with Wiener's concerns, as the current generation of AI agents and their human users together constitute social machines that can choose their own actions, while pursuing goals that are not necessarily set by their designers.

Most of the above examples are drawn from the area of online entertainment or commerce but in future such machines may not remain confined to those domains. As Turing warned in 1948, "the limited character of the machinery which has been used until recent times […] encouraged the belief that machinery was necessarily limited to extremely straight-forward, possibly even to repetitive, jobs". These systems are likely to be used in administration, education, and governance. Various proposals to use them for social regulations have been reviewed in (Cristianini and Scantamburlo 2019).

One important point is that there are agents whose actions are best understood in the light of their 'goals', as defined above, and this might help with their legal regulation. These are not the intended goals of the designers, nor the goals of the participants, but the emergent goals of the overall system, which may also drift as different participants are co-opted.

An important technical question for the future of the field is: how do we control such AI-enabled SMs? Setting aside the question of who has the right to do so, can the organisation running one such system really control its specific behaviour? And a related ethical question is: who is responsible for the behaviour of a SM? The users? The designers? The participants? The system itself?

We also need to face the important question of the effect of SMs on the individual participants and their autonomy and privacy, if SMs are—as it seems—capable of nudging and steering, and of accessing private information about their users. The Cambridge Analytica controversy about individual persuasion enabled by access to psychometric profiles of users, which were created based on their ordinary activities (Cadwalladr and Graham-Harrison 2018) is an example of a much wider issue. The fundamental point in all this is defining the relation between human users and AI-enabled SMs, a problem that is currently open from a technical, social, legal, and humanistic viewpoint.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Alexander J, Giesen B, Munch R, Smelser N (eds) (1987) The micro-macro link. University of California Press, Los Angeles

Beer S (2002) What is cybernetics? Kybernetes 31(2):209–219

Berners-Lee T, Fischetti M (1999) Weaving the web: the original design and ultimate destiny of the World Wide Web. Harper Collins, New York

Boden M (1990) Escaping from the Chinese Room. In: Boden M (ed) The philosophy of artificial intelligence. Oxford University Press, New York, pp 89–104

Börgers T (2015) An introduction to the theory of mechanism design. Oxford University Press, Oxford

Burr C, Cristianini N, Ladyman J (2018) An analysis and model of the interaction between intelligent software agents and human users. Mind Mach 28(4):735–774. https://doi.org/10.1007/s11023-018-9479-0

Burr C, Cristianini N (2019) Can machines read our minds? Mind Mach 29:461–494. https://doi.org/10.1007/s11023-019-09497-4

Cadwalladr C, Graham-Harrison E (2018) How Cambridge Analytica turned Facebook 'likes' into a lucrative political tool. The Guardian

Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356(6334):183–186

Cristianini N (2010) Are we there yet? Neural Netw 23(4):466–470

Cristianini N, Scantamburlo T (2019) Social machines for algorithmic regulation. Artificial Intell Soc 35:645–665

Covington P, Adams J, Sargin E (2016) Deep neural networks for youtube recommendations. RecSys's, Boston, pp 15–19

Difallah, D., Filatova, E., & Ipeirotis, P. (2018) Demographics and dynamics of mechanical turk workers. In: Proceedings of WSDM 2018: the eleventh ACM international conference on web search and data mining, Marina Del Rey, (WSDM 2018)

Facebook (2021) Facebook reports second quarter 2021 results. https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-Second-Quarter-2021-Results/default.aspx

Hadfield-Menell D, Milli S, Abbeel P, Russell SJ, Dragan A (2017) Inverse reward design. In: Advances in neural information processing systems

HLEG, ECAI (2019). A definition of Artificial Intelligence: main capabilities and scientific disciplines

Hofstaedter D, Dennett D (1981) The mind's I: fantasies and reflections on self and soul. Basic Books, New York

Jonauskaite D et al (2021) English colour terms carry gender and valence biases: a corpus study using word embeddings. PLoS ONE 16(6):0251559

Ladyman J, Wiesner K (2020) What is a complex system? Yale University Press, London

Maturana HR, Varela FJ (1991) Autopoiesis and cognition: the realization of the living. Springer Science & Business Media.

Mcfarland D, Bösser T (2002) Intelligent behavior in animals and robots. MIT Press, London

Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. Proc Natl Acad Sci 110(15):5802–5805

Kurtzweil R (2002) Locked in his Chinese room. In: Richards J (ed) Are we spiritual machines: Ray Kurzweil vs the critics of strong AI. Discovery Institute, Seattle, pp 128–171

Oard D, Kim J (1998) Implicit feedback for recommender systems. In: Proceedings of the AAAI Workshop on Recommender Systems, 81–83

Ricci F, Rokach L, Shapira B (2011) Introduction to recommender systems handbook. In: Ricci F et al (eds) Recommender systems handbook. Springer, Berlin, pp 1–35

Rosenblatt A (2018) When your boss is an algorithm, New York Times, https://www.nytimes.com/2018/10/12/opinion/sunday/uber-driver-life.html Accessed 17 October 2018

Rosenblueth A, Wiener W, Bigelow J (1943) Behavior, purpose and teleology. Philos Sci 10:18–24

Russell S, Norvig P (2002) Artificial intelligence: a modern approach, 2nd edn. Prentice Hall

Shadbolt N, O'Hara K, De Roure D, Hall W (2019) The theory and practice of social machines. Springer International Publishing, New York

Smart P, Shadbolt N (2014) Social machines. In: Khosrow-Pour M (ed) Encyclopedia of information science and technology. IGI Global, Hershey, pp 6855–6862

Smart P, Simperl E, Shadbolt N (2014) A taxonomic framework for social machines. In: Miorandi D, Maltese V, Rovatsos M, Nijholt A, Stewart J (eds) Social collective intelligence: combining the powers of humans and machines to build a smarter society. Springer, Cham, pp 51–85

Soper S (2018) Amazon's clever machines are moving from the warehouse to headquarters. Bloomberg, https://www.bloomberg.com/news/articles/2018-06-13/amazon-s-clever-machines-are-moving-from-the-warehouse-to-headquarters. Accessed 11 December 2018

Sutton A, Lansdall-Welfare T, Cristianini N (2018) Biased embeddings from wild data: Measuring, understanding and removing. In: International symposium on intelligent data analysis. Springer, Cham

Turing A (1948) Intelligent machinery, report for the national physics laboratory, http://www.turingarchive.org/browse.php/C/11. Accessed 10 January 2019

von Ahn L, Dabbish L (2004) Labeling images with a computer game. In: Proceedings of the SIGCHI conference on human factors in computing systems (CHI '04), pp 319–326 https://doi.org/10.1145/985692.985733

von Ahn L, Blum M, Langford J (2002) Telling humans and computers apart (automatically) or How Lazy Cryptographers do AI. CMU Tech Report. https://www.cs.cmu.edu/~mblum/research/pdf/tell.pdf. Accessed 23 October 2018

von Neumann J, Morgenstern O (1953) Theory of games and economic behaviour. Princeton University Press, Princeton

Wakabayashi D, Conger K (2018) Uber's self-driving cars are set to return in a downsized test, The New York Times, https://www.nytimes.com/2018/12/05/technology/uber-self-driving-cars.html. Accessed on 10 December 2018

Wiener N (1948) Cybernetics, or control and communication in the animal and the machine. The MIT Press, Cambridge

Wiener N (1954) The human use of human being. Da Capo Press, Boston

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

Zendle D, Bowden-Jones H (2019) Is excessive use of social media an addiction? BMJ 365:l2171. https://doi.org/10.1136/bmj.l2171

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.