AI & SOCIETY (2020) 35:519–532
https://doi.org/10.1007/s00146-019-00928-5

ORIGINAL ARTICLE

Behavioural artificial intelligence: an agenda for systematic empirical studies of artificial inference

Tore Pedersen[1,3] · **Christian Johansen**[2]

Received: 23 July 2019 / Accepted: 29 November 2019 / Published online: 18 December 2019

Abstract
Artificial intelligence (AI) receives attention in media as well as in academe and business. In media coverage and reporting, AI is predominantly described in contrasted terms, either as the ultimate solution to all human problems or the ultimate threat to all human existence. In academe, the focus of computer scientists is on developing systems that function, whereas philosophy scholars theorize about the implications of this functionality for human life. In the interface between technology and philosophy there is, however, one imperative aspect of AI yet to be articulated: *how do intelligent systems make inferences?* We use the overarching concept "Artificial Intelligent Behaviour" which would include both cognition/processing and judgment/behaviour. We argue that due to the complexity and opacity of *artificial inference,* one needs to initiate systematic empirical studies of artificial intelligent behavior similar to what has previously been done to study human cognition, judgment and decision making. This will provide valid knowledge, outside of what current computer science methods can offer, about the judgments and decisions made by intelligent systems. Moreover, outside academe—in the public as well as the private sector—expertise in epistemology, critical thinking and reasoning are crucial to ensure human oversight of the artificial intelligent judgments and decisions that are made, because only competent human insight into AI-inference processes will ensure accountability. Such insights require systematic studies of AI-behaviour founded on the natural sciences and philosophy, as well as the employment of methodologies from the cognitive and behavioral sciences.

Keywords Artificial intelligence · Artificial inference · Behavioral artificial intelligence · Artificial intelligent behaviour · Bias · Transparency · Accountability · Ethics

1 Introduction

Artificial intelligence (AI) can contribute significantly to the expansion of human judgment and decision making: AI learns and processes, while making inferences and drawing conclusions, similarly to what humans do, although with unsurpassed speed and capacity for handling complexity and volume. The strongest AI-algorithms are in fact developed based on knowledge about how the human brain functions and insights into AI-processes are limited by the same characteristical constraints that limit insights into the human brain:

- In the same way as humans may have difficulties in accounting for how they have processed the information that has lead them to make a judgment, AI is characterized by having the same type of "difficulties", in that advanced AI-algorithms are not immediately accessible and comprehensible, not even for the experts that pro-

Tore Pedersen was partially supported by the project Oslo Analytics funded by the IKTPLUSS program of the Norwegian Research Council. Christian Johansen was partially supported by the project IoTSec—Security in IoT for Smart Grids, with Number 248113/O70 part of the IKTPLUSS program funded by the Norwegian Research Council.

✉ Tore Pedersen
tore.pedersen@feh.mil.no

Christian Johansen
cristi@ifi.uio.no

[1] Center for Intelligence Studies, Norwegian Defence Intelligence School, Oslo, Norway

[2] Department of Technology Systems, University of Oslo, Oslo, Norway

[3] Department of Psychology, Bjørknes University College, Oslo, Norway

grammed them (Dahl 2018). The strongest AI-algorithms have already earned the status of being "black boxes",[1] in very much the same way as the processes of the human brain are inaccessible for oversight and comprehension. The reason is simply the enormous complexity of the AI machinery.

- Inaccessibility to oversight of complex AI-algorithms inevitably leads to lack of accountability, which is an imperative prerequisite for citizens' trust in parliament, public authorities, the justice system, or to trust that private enterprises conform to regulations or contracts. Accountability is a term that comprises important aspects such as the provenance of, access to, transparency of, and auditability of, algorithms and data. All these are important also for the reliability and validity of AI inference processes, which is a prerequisite for the assessments of AI accountability. But accountability is not studied sufficiently within the domain of Artificial Intelligence (Wachter et al. 2017) even though concerns about the relation between accountability and bias was raised almost 30 years ago (Dix 1992).

The general public has low awareness of how even simple technology is being used in democratic processes, policy- and decision-making. With an increased complexity in technology as a result of artificial intelligence, one is at risk of making judgments and decisions that do not conform to the requirements of accountability (Danaher 2016). When facing a technological complexity that is difficult or even impossible to comprehend, this may give rise to feeling disempowered. The satirical comedy "Little Britain" and the scene where "the computer says no"[2] may also be interpreted as showing how humans become enslaved when facing the "computer's" sovereign and opaque autonomy.

A new agenda The technological development within the domain of AI is predominantly fueled by needs for functionality and cost-efficiency, while other aspects, such as safety and security, ethics, transparency, accountability and judicial aspects, have not attracted the same interest, although these aspects have had much attention from theorists within arts and humanities, most predominantly stated as philosophical concerns about the possible negative implications of AI on humans and society (e.g., Binns 2018; Bostrom 2014; Dahl 2018; De Laat 2018; Pagallo 2018). Even if many have articulated **why** we need to gain more knowledge about AI and its opaque processes, no one has yet disclosed the **link** between AI-behaviour and AI-inference by describing **how** to study *AI Behaviour*. We argue in this paper that, even if complete transparency of AI processes is not achieved (Miller et al. 2017; Samek et al. 2017; Biran and Cotton 2017; Doshi-Velez and Kim 2017; Murdoch et al. 2019), for similar reasons as human cognitive processes are not immediately accessible from the 'outside', another type of transparency can nevertheless be achieved. By means of systematic studies of AI Behaviour, one could provide insightful evidence for the important aspects of AI, such as accountability and soundness of inference.

We will parallel later in the paper our proposed Artificial Intelligence Behavioural studies with Tversky and Kahneman's (1974) studies of human cognition and behaviour (and of the many ways in which people systematically deviate from a rational agent behaviour) which ultimately lead to novel insights into, until then, similarly 'opaque' cognitive processes. Eventually, this insight leads to the discovery of the relationship between heuristic processing mechanisms and biased judgments. Our starting point is the perspective of the theorists that have articulated the need for more knowledge about **why** one needs more insight into important aspects of AI output. We extend this perspective, elaborating on the **why** by comparing human cognition and learning to algorithmic processing and AI learning, and by introducing the concept of AI inference and AI behaviour to the discussion. As a result, we propose a novel agenda for **how** the theorists' articulated **why** can be achieved.

Consequently, there are two imperative aspects that are not yet well articulated:

1. The importance of studying *artificial intelligent behavior* and *artificial inference* systematically by means of empirical studies and experimental methods from behavioural and cognitive sciences, and
2. The need for broad scientific competence in order to assess whether or not artificial intelligent judgments and decisions are reliable, valid and inferentially sound.

The first aspect regards the need for research on AI-behavior and inference as a consequence of the artificial intelligence's autonomous and self-learning algorithms being opaque and thus inaccessible to human oversight and control of the validity of the output. The second perspective regards accountability in public and private affairs due to the fact that artificial intelligent judgments and decisions must ultimately be regarded as a *support* for human or institutional judgments and decisions. These aspects require a broad competence derived from several scientific disciplines, researching how artificial intelligent behavior can be studied systematically, as well as whether or not the artificial

[1] Knight, Will (14 March 2017). "DARPA is funding projects that will try to open up AI's black boxes". MIT Technology Review. https://www.technologyreview.com/s/603795/the-us-military-wants-its-autonomous-machines-to-explain-themselves/.

[2] Sample, Ian (5 November 2017). "Computer says no: why making AIs fair, accountable and transparent is crucial". the Guardian. https://www.theguardian.com/science/2017/nov/05/computer-says-no-why-making-ais-fair-accountable-and-transparent-is-crucial.

intelligent behavior conforms to established principles of reliability and validity in inferences. Soundness of inference and absence of bias—or at least awareness and reduction of bias—is a prerequisite for how much the processed knowledge can be trusted.

We thus propose a joint research endeavour—comprising both philosophical theorists from the arts and humanities, technology-developers from computer science, and empiricists from the social sciences, such as for example cognitive- and behavioral scientists—called *Behavioral Artificial Intelligence,* which would incorporate these perspectives. Behavioral Artificial Intelligence (BAI) would study the artificial inferences inherent in, and the manifested behaviour of, artificial intelligent systems in the same way as the social sciences have studied human cognition, inference and behaviour. We follow the traditions from other hybrid disciplines such as Behavioral Economics (e.g., Kahneman and Thaler 2006), Behavioral Transportation Research (e.g., Pedersen et al. 2011; Gärling et al. 2014) and Behavioural Computer Science (Pedersen et al. 2018). The key aspect in these disciplines is that they complement and update traditional fields such as economics, transportation research and computer science, by providing (descriptive) empirical data about actual human cognition and behaviour from the cognitive sciences and the behavioral sciences instead of relying on assumptions about (prescriptive) rational behaviour.

Of particular importance in BAI are:

1. the relation—similarities and differences—between human cognition and algorithmic processing;
2. the relation between human learning and algorithmic (machine) learning; and
3. the process of inferring knowledge from data, thus arriving at valid and reliable judgments, made by an artificial intelligence system compared to how humans make judgements.

In the following sections, we first elaborate on human cognition, learning and behaviour, and show how these are related to machine processing and output. Then we link reasoning and inference to important philosophical concepts, and show how human and machine inference must conform to sound principles of knowledge-generation, By thus disclosing the link between (AI-)inference and (AI-)behaviour we show that insights into AI-inference must be achieved with systematic studies of AI-behaviour, in the same way as insight into human inference have been achieved with systematic studies of human behaviour.

2 Human cognition and artificial intelligence processes

Artificial intelligence is now widely covered in the media, generating feelings ranging from fascination to fear (LeCun et al. 2015; Jordan and Mitchell 2015; Parkes and Wellman 2015; Boden 2016; Biamonte et al. 2017; Frégnac 2017; Ramprasad et al. 2017; Brundage et al. 2018). **But what *is* artificial intelligence?** A brief glance into the history of AI shows that Alan Turing, in his endeavour to construct "the mechanical brain" in the 1950s, was a prominent actor in the early development of artificial intelligence. Later developments included the works of the 1978-recipient of the Nobel Prize in Economics, Herbert A. Simon (Newell et al. 1958, 1972, 1976; Simon 1979, 1996) as well as the 1971-recipient of the ACM Turing Award, John McCarthy (McCarthy 1987; McCarthy and Hayes 1969). Although early AI-endeavours were predominantly in the symbolic tradition, based on a rational actor paradigm including forms of logic and reasoning systems, whereas later endeavours are non-symbolic and more influenced by empirical neuroscience (cfr Dix 2016),[3] their works nevertheless show that a significant part of the processes in AI are inspired by contemporary-times' knowledge about human cognition and how the human brain functions.

AI and machine learning methods such as neural networks and reinforcement learning (Sutton and Barto 2018) are examples of current algorithms directly inspired by human cognition, whereas evolutionary computation and genetic programming (Holland 1992; Bäck et al. 1997; Mitchell 1998; Coello et al. 2007) are inspired from human biology and evolution. Whereas the symbolic AI-tradition is related to the cognitive System 2 and thus based on prescriptive rational agency, the non-symbolic empirical tradition is related to the cognitive System 1 and thus more descriptive, based on empirical evidence for (also non-rational) behaviour. One of the most popular concepts today is *deep neural networks* (DNN) (e.g., Demuth et al. 2014), which is even more similar to how the human brain functions. In DNN, artificial neurons are connected in a network where each single neuron communicates with a great number of other neurons, in very much the same way as the neurons in the human brain do. The network consists of several layers of neurons. One "input" layer is responsible for taking in a problem instance, such as an image, and one "output" layer is responsible for producing the response, like yes/no whether a cat is present in the image. Several other "hidden" layers of neurons are involved in the processing, hence the

[3] The EPSRC 'Human-Like Computing' initiative aims to bridge this 'gap' between 'symbolic'/'rational' and 'neural'/'empirical' AI. See: http://hlc.doc.ic.ac.uk/.

concept "deep" and each single neuron encodes a specific aspect of the "problem" the DNN is designed for. In our problem of recognizing cats in images some neuron will encode the fact that cats have a visible tail (or something even more seemingly insignificant than that). AI techniques are generally built in two stages: first *a learning stage* where the DNN is trained on a dataset where the correct response is known for every input; and then *an operational stage* where the algorithm is used in the wild; given a new input it produces a decision or action (such as changing the picture or logging some inferred information about the picture). Sometimes these stages are combined, and an algorithm constantly learns if it has a mechanism to obtain some form of feedback on the responses that it produces. Even though there are differences between DNNs and the human brain, for example, that DNNs are dependent on massive volumes of data in order to learn, whereas the human brain may be able to learn from being exposed to only one single instance of a phenomenon (i.e., data),[4] the similarities are nevertheless striking—and for both, the learning processes will always result in some type of behavioural output.

Simple algorithms versus artificial intelligence Machines can be "simple", in the way that they are given predefined instructions—that is, algorithms—to perform a task and never deviate from these instructions when processing data, e.g. when locating specific information in a dataset or making a judgment about some attributes of the dataset (Minsky 1967; Knuth 1973; Dennett 1995). In simple machines, the processing is transparent in the sense that it is possible, at least to some extent,[5] for humans both to understand the process and to have access to what the machine actually does with the data it processes. Algorithms when executed by a machine can be seen as the "behavior" of the machine (e.g., the returned results, the commands sent to some actuators, etc.) terminology, which appears in various specific areas of computer science research (Hennessy 1988; Bergstra et al. 2001; Hennessy and Rathke 2004; Ancona et al. 2016). However, Artificial Intelligence (Russell and Norvig 2016) is an example of a "complex" machine which is programmed (with algorithms) in a way that allows it to interact with its surroundings (in the form of data continuously fed in) and as a result of this interaction is allowed to make new algorithms and adjust existing algorithms after having "learned something new" from the surroundings. The adjusted internal execution of AI thus can be seen as adapted new behavior that AI learns continuously. Whereas the processes in a simple machine are transparent, the processes in intelligent systems are, as a consequence, opaque and thus not immediately accessible to human oversight, supervision and audit (Russell 1997; Rahwan and Simari 2009).

To have an intuition about the AI learning process, consider for example how the brain of a child develops through learning, beginning with birth and continuing by interaction with other humans as well as other parts of the child's surrounding environment. The child's brain is being "wired" in a particular way as a result of learning through such interactions, which we call experience. This can be compared to the way artificial intelligent systems functions: during the learning phase the neural network will be "wired" in a particular way as a result of interacting with the training data—thus, the training data represents the artificial intelligent system's experiences. But what will happen if one allows an artificial intelligent network unlimited possibilities to learn through unlimited interaction with voluminous data sets over time and, moreover, the network is allowed to autonomously develop its "deep" neural network outside of human control (something which already is done by significant AI-actors such as IBM and MIT)? What will such an artificial intelligent system be like, compared to the human brain? Will the AI system develop a personality, abilities, recognizable traits? What sorts of preferences will the system have and what types of inference-criteria will be the basis for its generation of valid and reliable knowledge? Will it always make objective, accurate and "just" judgments, or could it be biased in the same way as a human can be biased? In other words: how does AI infer and what behaviour does it exhibit?

To understand how artificial intelligent systems make judgments and provide decision-support for human and institutional decision-making in the context of identifiable accountability, one should adapt methodologies from empirical studies on human personality as well as on human judgment and decision making, from the cognitive sciences and the social and behavioral sciences. This would imply using cross-sectional, longitudinal or experimental designs. Cross-sectional designs would allow to compare a broad range of AI systems to each other with regard to characteristics such as 'personality and preference' and the validity and reliability of AI 'judgments'. Longitudinal designs would allow us to monitor a single AI system's 'personality' and 'judgment' over time, providing information about possible fluctuations. Experimental studies would allow manipulations to help detect whether some types of input (or interactions) would affect 'personality' and 'judgment' more than others. True experimental design (also known as randomized controlled trials: RCT) would even allow the study of potential biases in AI systems as well as the causes of the biases (i.e., by means of manipulations), something which would provide valid and reliable answers to the concerns raised about

[4] Cfr stimulus—response and classical conditioning.

[5] Even 'simple' codes can sometimes be difficult to fully comprehend, e.g., probabilistic programs (Gordon et al. 2014; Katoen et al. 2015) or concurrent programs (Andrews and Schneider 1983; Ben-Ari 2006; Dijkstra 1965).

machine learning biases, both by academe (Baeza-Yates 2018; Crabtree et al. 2019; Dix 1992, 2018; Dwork 2011; Monroe 2018; Zemel et al. 2013) and regulatory authorities (USACM 2017).[6][7] These concerns are explicitly raised, but yet no reliable answers have been provided, something which strongly signals the need for scientists to respond to the concern by exploring methodology in novel ways to arrive at the answers.

Extensions of Man McLuhan describes in *Understanding Media; The Extensions of Man* (McLuhan 1964) the difference between *information* per se and the *specific medium* in which the informational content is transmitted, e.g., in writing, orally, visually, tactile. An important point made by McLuhan is that a medium is an extension of human traits and behaviour, and thus an extension of humans as such, e.g.: the written word is an extension of thought and speech; tools and utensils are extensions of fingers and arms, clothing is an extension of the skin. Nowadays, many types of technology can be seen as even more advanced extensions of humans. We can easily perceive artificial intelligence as an extension of human cognition, memory—and inference.

Another aspect in McLuhan's analysis is that each and every medium has an inherent tendency to shift human attention away from the informational content that the medium is a vehicle for, and over to the medium itself, something which leads to the perception of the medium *as* the actual informational content, instead of being a vehicle for transmitting and communicating the informational content. Today, AI and the opacity of its "black box" processes may, in the same way as McLuhan perceived traditional media, easily become a medium that draws attention away from the judgments it actually carries out.

Machine learning algorithms (which we interchangeably call AI in this paper) are particularly well suited to quickly elicit correct information (probabilistically, i.e., to some degree of error) from a voluminous dataset, also if the data consists of information that is not structured. Natural language, or images, as opposed to data that is presented as numbers in rows and columns, are considered unstructured data. IBM Watson is an example of a machine capable of finding a correct answer from a voluminous dataset, provided that there actually exists a correct answer (Ferrucci 2011; Ferrucci et al. 2013). IBM Watson cannot, however, look for clues to the answers "outside" the recorded datasets in the same ways as humans can—something which of course is a limiting factor for these types of intelligent systems. Because a machine such as IBM Watson can process voluminous information very fast, as opposed to a human who has limited capacity for processing and who is also considerably slower, the machine is an invaluable helper when large amounts of information need fast processing.

In *Technopoly: The Surrender of Culture to Technology* Neil Postman (Postman 1992) made an analysis of technology that in many ways resemble McLuhan's thoughts: all types of traditional tools and technologies have inherent traits that make them particularly good at carrying out specific tasks, such as the axe is good at cleaving, or the knife is good at cutting. Gibson (1979) adds momentum to McLuhan and Postman's analyses in a behavioural sense by showing that external objects are perceived by humans not only in the sense of having a particular shape or being placed in spatial distances from each other but also by their 'affordances'. According to Gibson's Affordance Theory, objects are perceived as having **potential for action**—that is, it is the human perception of objects that drives (or prompts) human action. Thus, if a knife is good at cutting, it is because humans "allow" the knife to exhibit its inherent traits by being vehicles for the knife's potential. In the same way, advanced technologies also have inherent traits and tendencies which make them good at carrying out specific tasks. Thus, a tendency emerges which makes the various types of tools and technologies "promote" the particular types of tasks that each tool and each technology is good at performing. However, whereas the inherent traits of simple tools are easily decoded, the same traits are not easily decoded—nor even discovered—in advanced technology. Thereby we forget—or, more correctly, we do not understand—AI's inherent traits and accompanying tendencies to promote specific processes that lead to specific outcomes. For example, when AI learns by interacting with a limited dataset that represents only a fragment of the social world, inferences about the larger social world may be drawn from a non-representative dataset, much in the same way as humans are also prone to do if not 'corrected'. Thus, one specific implication of an AI 'trait' is that AI would exhibit the same biases as humans, only extremely more 'efficiently'. This could lead to a substantial increase in biased judgments as a result of the powerful AI extension of human erroneous reasoning. An example of this powerful effect, albeit from traditional technology that is more familiar, is the grading of high school exams and the subsequent acceptance or rejection of students into universities or into the labor force. Today, we accept this technology as 'a given' (or even 'a natural') expression of human knowledge and skills when, in fact, the most prominent feature of exam grading is administrative simplification with the aim of ranking and sorting humans

[6] New York City Council (2018). A Local Law in relation to automated decision systems used by agencies. http://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0.

[7] EU Parliament (2016). EU Framework on algorithmic accountability and transparency. http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+WQ+E-2016-007674+0+DOC+XML+V0//EN.

for some other end-purpose and merely taking the grade as a face-value 'token' for knowledge.

Adapting research methods from the cognitive sciences and the behavioural sciences to artificial inference and behaviour would help gather empirical data that will provide increased knowledge and insight into whether AI systematically exhibit specific and recognizable traits and preferences in the same way as humans do, and whether AI exhibit systematic tendencies to "behave" in a certain way in terms of having a tendency to employ specific types of inferences or to produce specific types of output (e.g., fallacies in formal reasoning or whether judgments and decisions are biased due to data not being representative of the 'true' population). It would, however, be interesting to see if future AI would be able to avoid being 'blind' to its own biases—that is, if AI would be able to somehow 'transcend' the ubiquitous human tendency to find justifications for own biases.

Whereas the research traditions in the arts, the humanities, and the social sciences, most notably Philosophy, have been concerned with the implications of technology for human life and human activities, the technological research traditions themselves have been more occupied with developing new technology with the aims that technology should solve existing problems such as replacing the human in the workplace. Thus, the most important aspect for the technological sciences is the development of functionality. As a consequence philosophy has been focusing on ethical aspects of technology and has been concerned with the potential threats to humanity posed by technology if humans lose their control over technology (Bostrom 2014). The "division" of these two research fields is an example of a compartmentalization when the fields would, in fact, benefit from being multidisciplinary. This is why it is necessary to embrace a more holistic approach to AI—an approach that incorporates both empirical evidence about both the technological development of AI, and implications for the individual and societal domains in which AI is employed, including actual artificial inference and artificial intelligent behaviour. Regardless of whether it is a human or a machine that observes, make judgments or inferences, and concludes—individually or in interaction with its environment—both humans and artificial intelligent systems must nevertheless adhere to the same principles when generating knowledge that is supposed to be valid and reliable. In other words, they must all build their inferences on sound reasoning and representative empirical data.

3 Artificial intelligence and cognitive bias

Two modes of thinking The cognitive and behavioral sciences have through decades provided substantial empirical evidence on how humans think, make judgments, assess, choose, and make decisions. The empirical evidence has resulted in theories and models that show a distinction between two typical modes of cognitive processing—that is, thinking—labeled System 1 and System 2 (Kahneman 2003; Sloman 2002). Whereas System 1 is a non-deliberate and automatic mode of thinking that is carried out by our brain while we are not consciously aware that we are making a judgment or an assessment, System 2 is the deliberate and non-automatic mode of thinking that we carry out while being absolutely consciously aware that we are making a judgment or an assessment. Cognition carried out by System 1 is commonly labeled *intuitive thinking*, whereas cognition carried out by System 2 is labeled *analytical thinking*. When we make "formal" judgments and decisions in situations where we are aware of what we are doing and informed in the sense that we have relevant information of high quality and sufficient time to make a judgment, we are carrying out a rational and controlled judgment of a case by means of System 2.

In the context of the continuously processing and always-learning artificial neural network, the human System 1 could be seen as paralleling the automatic and fast response that the network provides when a new problem needs to be solved, whereas the human System 2 could be seen as paralleling the AI's feedback loops and reinforcement learning that identifies when the answer is not correct, and recalibrates the network so that it is more likely to produce a more accurate judgment and thus a more correct solution to a future similar problem. However, this tentative analogy is not clear-cut, as there are also similarities between the human analytical System 2 and the traditional 'symbolic' AI-endeavours. Another analogy is between the human analytical System 2 and the traditional symbolic AI endeavours based on logics, which we see coming back into fashion to explain AI (e.g., Bottou 2014; Vardi 2018).

Intuitive thinking: heuristics as bias-generators On the other hand, many judgments and decisions that we make in everyday situations are often carried out without us being consciously aware of the fact that we are making a judgment or a decision. This is an example of a normal cognitive processing mode that is appropriate in many everyday situations. Our brain automatically filters out stimuli that are not relevant to the particular judgment that we are going to make *right now*. The process is going on without our conscious awareness, and it usually functions well. As seen from an evolutionary perspective, this intuitive reasoning-mode was highly appropriate in a less complex environment than ours; today, however, heuristics may lead to a biased output under conditions (and in environments) that are characterized by uncertainty. Thus, the appropriateness of the reasoning-mode depends on the context: when we intuitively possess the knowledge and the skills needed to arrive at a correct judgment, fast heuristic reasoning works well. However, our

brain has a tendency to automatically activate System 1 also in situations where we should instead have activated System 2 (e.g., in situations where we have too little knowledge or too little experience). This can easily lead to an incorrect judgment, usually called a *cognitive bias* (Tversky and Kahneman 1974), caused by psychological mechanisms that are activated when we make judgments under uncertain conditions (e.g., when information is incomplete or too complex, or when the time-frame at our disposal is limited). These psychological mechanisms are labeled *heuristics* and they function as mental shortcuts activated without conscious awareness (Gilovich et al. 2002; Tversky and Kahneman 1974).

We mention two widespread heuristics (Kahneman 2003):

- *The accessibility/availability heuristic* (Tversky and Kahneman 1973) prompts us to infer (non-consciously) that information easily accessible in memory represents something that is also common or frequently occurring. For example, since it is easier to recall words beginning with the letter R than words where R is the third letter, people may wrongfully infer that words beginning with R are more frequent, just because the former is easier to recall (Tversky and Kahneman 1974). Similarly, we may make a biased judgment about a phenomenon simply because there is high availability only to some instances in memory and not to others.
- *The representativeness heuristic* (Kahneman and Tversky 1972) prompts us to make an instant judgment of an unfamiliar phenomenon based on a perceived similarity to a familiar phenomenon, even if this similarity is only superficial. For example, the colour of someone's skin, the way they dress, or the way they talk, are characteristics that, when they occur in combination with demographic features such as age, gender, or occupation, may prompt us to infer that specific individual characteristics and specific demographic features are strongly related. Thus, we may make a biased judgment about a person simply because we make an incorrect categorization due to an 'overactive' System 1.

If we are given feedback on a judgment we make, we will probably experience learning (although learning may also occur by the association as a result of classical conditioning), and a future judgment may become more correct. But if System 1 does not receive feedback, or receives false feedback (e.g., not sufficiently, or wrongly "calibrated" to reality) we will not learn that the judgment is incorrect, and the incorrect judgment may sustain and be applied also in future situations. Interestingly, reinforcement learning in AI uses the same feedback methods (Sutton and Barto 2018). The science of human thinking and learning have long traditions and convincing empirical evidence from longitudinal as well as experimental studies, both controlled in laboratory and taking place as naturalistic studies in the field. We argue that one must carry out empirical studies also of *artificial inference and learning*.

The two modes of thinking are similar to the way we learn. For example, children are given little or no formal education when they are learning their native language. Moreover, they are not consciously aware that they are learning a language. Thus, they learn the language intuitively, continuously interacting with their surroundings. As Burling (2005) notes, pertaining to learning, the most interesting part of language lies in the interplay between the production and the reception of language. This mode of learning has much in common with intuitive thinking. In the same way, we learn as adults to make judgements about the world surrounding us and the everyday-situations. This intuitive learning process—also labeled implicit learning without conscious awareness (Frensch and Runger 2003)—is strongly related to intuitive judgments prompted by System 1. On the other hand, we may, of course, be consciously aware that we are learning and that we are in fact taking part in a learning process. When adults sign up for a course to learn a foreign language they are consciously aware of the learning. This learning process has much in common with the judgments we make when we are employing analytical thinking in System 2. In addition to having conscious awareness of the process, we will also receive explicit feedback about whether we have learned what we have been expected to learn. Thus, we learn to use the foreign language correctly and we avoid making mistakes.

Human aspects in the machine It has generally been assumed that a machine is objective and infallible compared to humans, but this has been widely questioned already since Alan Turing, and more recently also in legislations (EU, 2017; New York City Council, 2018; ACM US Public Policy Council, 2017). Most notably, a serious concern was raised in 2018, in a paper jointly signed by authors from prominent US and UK universities and research institutions, about the malicious use of artificial intelligence (Brundage et al. 2018). The common concern of these initiatives is the discovery that intelligent algorithms make incorrect judgments, something which has been known to lead to unethical and "unfair" decisions. For example, algorithms discriminate people (e.g., based on ethnicity, gender, or age) in automated decision-processes, such as in financial matters, in medical diagnostics, and in law enforcement.

- One explanation to why such incorrect judgments are made is that an algorithm is always initially programmed by a human; it is always a programmer that creates the initial algorithm that the machine uses when carrying out its tasks. A human's incorrect judgments (biases) may thereby be transferred to the machine algorithm,

something which implies that the machine may exhibit the same incorrect judgments (biases) as the human that programmed it would have done.
- Another explanation, believed to be prevalent, is that the machine learns about the problem from a limited data set that does not necessarily represent all the aspects relevant for the problem. To put it another way, the algorithm learns from a limited sample that does not represent the larger population (i.e., the data is biased).
- Moreover, even if the machine makes a judgment that is correct in a formal and logical sense, the same judgment may be deemed inappropriate or unacceptable in an ethical sense, depending on the context; even if generalizing from demographic characteristics to individuals may be appropriate when screening for disease in order to provide health-care, it is unacceptable to reject job-applicants based on the same generalizations.

Thus, both the transfer of biases from programmers to machines and machine-generated biases, imply that machines are not necessarily more infallible than humans, even if they are substantially faster and can process larger volumes of data. In fact, both humans and algorithms are fallible, but must nevertheless do their best to conform to principles of logic and reason and sound inference, when imposing a judgment *on* something in the social world and when learning something *from* the social world.

The systematic study of AI-bias The aim of experimental-empirical behavioural research in the domain of psychology—more specifically, judgment, decision making, and **inference**—is to study whether humans exhibit general and universal tendencies to think, make judgments, decide, choose or behave in a certain way. For example, Tversky and Kahneman, as well as contemporary and recent researchers, studied, at an aggregate (group) level, the individual human tendency to non-consciously activate heuristic thinking under conditions of uncertainty and thereby exhibit distinct types of biases in their judgments. If such tendencies is to be elicited in a scientifically valid way, one needs to select random samples of participants from a known population and carry out true experiments—that is, one or more variables need to be manipulated, the participants must be randomly assigned to either an experimental group or a control group, and all other conditions (except for the manipulated variable) must be identical for both groups. Analyses are then carried out using inferential statistics, either as regression or as analysis of variance, and results will show whether the sample at group level exhibits (or not) the hypothesized tendency. If the tendency exists, it is shown as a *mean* effect at the group level. After establishing the overall effect at the group level, the next step is to recruit new samples and analyse whether there are systematic *individual differences* in the tendency to exhibit the bias. For example, whether participants with some specific demographic characteristics (or specific personality characteristics, or specific cognitive "styles" of processing information), are more prone to exhibit the bias than others who do not share the same characteristics. Thus, this type of research starts with establishing the effect at group level and is pursued further by identifying differences at the individual level.

We argue that one needs to study AI-bias similarly. In order to systematically study AI-bias one must design and carry out empirical (preferably experimental) studies with AI-participants (i.e., systems or variations of their configurations or training data) in the same way as one has previously done with human participants. If one wants to gain knowledge about the potential general and universal tendency for AI to exhibit bias (both the biases that are already known, as well as hypothesized biases), one needs to study the phenomenon on an aggregate (group) level with a number of AI-systems as "participants". This way, one can detect a *mean* effect of the AI's tendency to exhibit a specific bias. A next step would be to look for "individual" differences at the individual AI-level, such as, whether AI-systems from one provider are more or less prone to exhibit a bias than AI-systems from another provider. There are now enough different AI systems for the same task, e.g. see the large AI competitions such as Kaggle competitions[8] or the traditional RoboCup events (Osawa et al. 1996). Even from a single AI system one can create a group of multiple diverse AI systems on which to study bias by training the AI system on different training data sets. Even on the same data set an AI system can be configured in multiple ways, all of these forming the group that we want to study for biases.

4 Inference: human *and* artificial

When crossing the border between the early traditional 'symbolic' AI-endeavours and the neuroscience-inspired 'DNNs' in later AI-endeavours—or when bridging this gap, such as in the initiative of the EPSRC Project 'Human-Like Computing'[9]—important questions arise regarding future AI's ability to adhere to the philosophical underpinnings of reasoning. Would future AI understand the difference between different forms of 'being', and would it be able to distinguish between different ways of acquiring knowledge? Would future AI distinguish between different approaches to understanding 'universal' versus 'unique' realities, and would it distinguish between different types of 'truth'? These reasoning-related philosophical aspects that all need to be adhered to regardless of whether the agent is human or an

[8] https://www.kaggle.com/competitions.

[9] http://hlc.doc.ic.ac.uk/.

AI-system, are all relevant to the study of AI-behaviour and the link between AI-inference and AI-behaviour.

Ways of knowing In philosophy and metaphysics, two concepts are central to understanding what it means that something *exists* and how we can *know* something about that which exists. The concept regarding various forms of existence is known as *ontology* (Coffey 1914) and the methods employed for acquiring knowledge about existence is known as *epistemology* (Coffey 1917). As an example, our physical world can be understood as one type of existence, whereas our social world can be understood as another—different—type of existence. A physical object, say, a chair has one type of existence in the physical world, but may have another—or an additional—type of existence in the social world. In the physical world we are concerned about the chair's characteristics as a physical object, whereas in the social world, we are concerned about the *meaning* and significance of the chair—that is, the role that the chair has been *given* in the social world that it occupies.

If one is interested in distinguishing between different types of existence, one concerns oneself about ontology as a field of study. Being able to distinguish between various types of existence is valuable both in occupational life as well as in private life. There are different types of methods to acquire knowledge about the different types of existence. If one wants to know something about Santa Claus's physical existence as an entity actually living at the North Pole, this would require one type of investigative method (e.g., look for traces of Santa on places where he is claimed to live), whereas wanting to know something about Santa as a collective representation in children's minds would require a different type of investigative approach (e.g., interviews with children in order to gain knowledge about the characteristics of this collective representation) (Dylan et al. 2017). Different ontologies as well as the different epistemological approaches in each of the ontologies can provide valuable knowledge, but it is nevertheless two different types of knowledge that we acquire with different methods that yield different meanings.

If one wants to make AI systems find out something about a social phenomenon or to make a judgment or a decision about such a phenomenon, one should investigate which type of ontology the AI-system is employing and whether the AI-system's investigative methods are suited for acquiring knowledge about the phenomenon. Since metaphysics and philosophy of science (i.e., ontology and epistemology) are difficult for most humans to understand, how can we make sure that AI-systems will be able to handle these concepts and provide valid and reliable knowledge about complex phenomena in the social world? In the same way as humans are prone to make mistakes about the ontology and corresponding epistemology they base their knowledge on, AI output could also easily become biased if AI—or a human, when programming or feeding data to the AI—chooses the 'wrong' ontology or the 'wrong' epistemology. An example of this is Google Flu Trends (GFT) (Lazer et al. 2014), an intelligent system which was developed to predict outbreaks and spread of Influenza. GFT was initially fed with the officially recorded prevalence of Influenza (the condition is identified by physicians and then registered in national health databases) and simultaneously fueled with data from the Google Search Engine (where people typically search for symptoms related to Influenza). Due to being calibrated to both the official records and people's searches for symptoms, GFT could initially provide accurate predictions of the future outbreak and spread of Influenza. However, the accuracy of these initial predictions decreased; GFT began to predict a substantially higher rate of outbreak and spread of the condition than what was the actual outbreak and spread. This happened because GFT was not able to distinguish between people's search activities when they unknowingly had a common cold and were concerned about their health without actually having Influenza, and their search activities when they did, in fact, have Influenza. Since search activities when people are concerned about their health are quite similar to the search activities carried out when they do in fact have the condition, it is not possible, neither for machines nor for humans, to make accurate predictions based on search activity alone, because 'searching for Flu' and 'actual Flu' are two different ontologies.

The particular and the general In addition to the existence of different types of methods for acquiring knowledge, it is also important to be clear about whether one wants to acquire knowledge about something that is specific, local and unique, or to acquire knowledge about something that is general, global and universal. The former is known as *idiographic* knowledge, whereas the latter is known as *nomothetic* knowledge (e.g., Robinson 2011; cfr Hurlburt and Knapp 2006). If one wants to acquire knowledge about a specific case one can conduct case-studies of the phenomenon or do in-depth interviews with the person in which the phenomenon resides, or even ask other people about the phenomenon that resides in that specific person. If we are interested in gaining knowledge about how one specific person describes the driving forces of his or her life—that person's life-narrative—we will, of course, gain valuable knowledge about how this person *understands* his or her life and ascribes or interprets meaning into it. This type of investigative approaches—generally known to lead to idiographic knowledge—would provide valuable insight into a specific, local and unique phenomenon, but it does not warrant the generalization of the phenomenon to also be present in other persons when we have actually not studied other persons—that is, other specific cases. To ensure sound reasoning when making inferences about individual cases

based on group statistics, the employment of System 2—analytic thinking—is crucial.

On the other hand, if one wants to know something about whether a phenomenon—say, a personal trait, a preference, a taste, a tendency—is shared by many people, one should try to study the phenomenon broadly and make sure that the sample of the study consists of many people drawn from a larger population so that the sample is representative. If one is interested in knowing, for example, if pre-school children prefer chocolate ice-cream over vanilla it is, of course, important to study a representative sample of pre-school children. One might find out that, in general, 60 percent of the girls prefer one flavor of ice-cream over the other and that boys are more concerned about eating an ice-cream and care less about whether it tastes chocolate or vanilla. This type of knowledge—generally known as nomothetic knowledge—would provide valuable insight into the distribution of preferences in a larger group or a larger population, something that is important for many reasons, but it would not help with understanding whether a new and unknown member of that same population would share the same preference or exhibit the same tendency. The knowledge is valuable for understanding tendencies on *group level* but does not automatically help us know whether we should serve chocolate ice-cream or vanilla ice-cream when a colleague is accompanied by her daughter to our office.

Now, let us ask ourselves: how would an AI-system go about to *understand the meaning* of a social phenomenon that exists in living people—such as a person's life-narrative—and how would an AI-system go about to decide what it takes to ensure that a sample is representative of a larger population? Is it even possible that an AI-system would be able to relate to concepts such as idiographic and nomothetic knowledge? The only way to answer these questions is to study AI judgment, inference and behaviour in the same way as we have been studying human judgment, inference and behaviour.

Three modes of inference The questions posed in the previous paragraph is closely related to the act of making an *inference*. If one observes a specific phenomenon during similar occasions and under identical circumstances, one may rightfully claim that the phenomenon will occur also on the next occasion, provided the circumstances are identical. The term for this type of inference is *inductive reasoning* (e.g., Anderson 1948; Bacon and Montagu 1857). After many observations of the same phenomenon we make a rule about the phenomenon: it will occur again if the circumstances are the same as in previous observations. A key term in the previous sentence was *many*. How many consecutive observations does it take to be convinced that the phenomenon will occur again the next time? The answer is that it depends on several things and that we can never be absolutely sure. Thus, probability is another key concept in inductive reasoning. In terms of meteorology, making a prediction that the gulf-stream—a part of the atlantic ocean current that brings warm water from the south to the north—will still be present tomorrow morning is a straightforward prediction to the extent that the gulf stream is a stable phenomenon. But in the case of non-stable phenomena or possible disruptions of stable phenomena, predictions are more difficult to make. Does the phenomenon currently seem to be somewhere on an upward trend? If so, we will often tend to extrapolate and predict that the ongoing trend will continue, although disruptions may occur and generate unforeseen surprises. To search for missing evidence and to know when induction might fail is difficult. Think for example of stock traders acting as if the current trend will continue forever.

How would an AI-system go about to determine how many observations it would take to engage in inductive reasoning and make a valid inference about the continuous future occurrence of a current phenomenon? Even if humans are fallible in making inductive inferences, an AI-system does not have the same opportunity as humans to be aware of data that is outside of 'reach' or to seek advice on this, so how would it keep its reasoning on track? How would AI understand that a disruptive event that is not directly related to the data could render further induction invalid? It seems that one needs to look for phenomena 'outside' the dataset, and even to look at other societal contexts, in order to infer whether further induction is appropriate or if it should be abandoned.

On the other hand, if you draw a conclusion—that is, if you *infer*—on the basis of a set of premises that are *given* (albeit not necessarily true) and thus are obligated to adhere to these premises, you are conducting reasoning strictly within the domain of formal logic. This type of reasoning is known as *deductive reasoning* (Kneale 1945) and is different from inductive reasoning in that the premises do not need to be true in terms of corresponding to observations. Correct reasoning in this inference-mode follows from formal logic only, so it is entirely possible for AI-systems to adhere to this type of reasoning, given that the premises in the formal arguments are recognizable to the system. However, there are not many things in the social world of human affairs that fit easily into such a structure of formal logic. It is, of course, possible to code instances in the social world and fit them into the structure of formal logic, so that AI-systems may be able to help us to make valid inferences, particularly if there are many premises in a structure of argument. If so, AI-systems are much better than humans to process a large number of informational components. Thus, this reasoning mode probably represents the potential of traditional AI at its best, but it needs human coding of instances before the AI-processing and it needs human oversight of AI-output. It is questionable whether an AI-system that only processes its inferences in deduction-mode can be called intelligent.

This difference between induction and deduction indicates that there are more than one way to understand the concept of truth—or to decide what it takes to call something *true*. One definition of truth is that a claim about, or a description of, something is true if the claim or the description corresponds with the "something" such as it actually is in the real world. If we make a claim about an instance in the real world, for example, that a specific apple in a specific fruit bowl on a specific table is green, and can also observe that this specific apple is, in fact, green, then our claim or description is true. The essence of such a definition of truth is *correspondence*: if our claim corresponds with things as they actually are, then our claim is true. Thus, this definition is generally known as *the correspondence principle of truth* (e.g., Kirkham and Kirkham 1992). The correspondence principle is closely related to inductive reasoning in that it presupposes—or at least favors—observations of instances so that they can be compared to what is claimed about them.

For deductive reasoning, however, it makes less sense to talk in relation to truth-as-correspondence. Deductive reasoning does not relate to an "external" observable reality in the same way as inductive reasoning does, but rather to an "internal" relation between the premises of an argument—that is, between the premises and the conclusion. Thus, deductive reasoning does not rely on empirical, observable data, in order to be valid; it concerns itself only about the internal consistency and coherence between the premises and the conclusion in an argument. This is known as the coherence theory of truth, or *truth-as-coherence* (Rescher 1973). If the whole argument is consistently coherent internally—regardless of whether the premises are "correspondence-true" or not—the argument is *true* in the sense that it is valid.

The value of the deductive inference mode and the coherence theory of truth is that it helps us reason in a way that follows formal logic. If we adhere to these principles, we can trust our reasoning more than if we do not adhere to them. Even if we have separated inductive reasoning from deductive reasoning, and even if we have separated truth-as-correspondence and truth-as-coherence, generating knowledge about the social world will more often than not employ both inductive and deductive reasoning. This process is generally known as the *hypothetico deductive method* (e.g., Popper 1972), first proposed by Christiaan Huygens in the 17th century,[10] in which one adheres to the principles of formal logic—to ensure consistency, coherence and sound reasoning—while at the same time observe whether the premises in the arguments, in fact, correspond with the same instances as they are observable or manifest in the external (social) world. Although science progresses more in the sense of shifting from 'old' to new paradigms when new evidence suggests that older paradigms must be discarded (Kuhn 1962) than by continuous Popperian falsification, many types of empirical research are carried out by means of the hypothetico deductive method.

One could claim that modern AI-systems have adopted more the correspondence-principle of truth, employing more an inductive reasoning where their training data comprises the many observations about the problem at hand. Reinforcement-style of AI would constantly add more observations of true or false instances of the problem. However, AI-systems are programs, and as such employ formal logic in their code through the programming languages and algorithms, and more recently also through the mathematical theories behind. Thus we would say that AI, more than other traditional IT-systems which are only logic-based and deductive, combine the two truth-principles, and maybe adhere more to the Popperian approach than other systems. However, we feel that this combination has not been sufficiently recognized nor harnessed. We would thus propose more work in this direction, combined with understandings coming from the social sciences, e.g., on the lines drawn by the EPSRC Project 'Human-Like Computing' and aiming at offering AI systems abductive-reasoning abilities.

Abductive reasoning Many instances in the social world, however, do not lend themselves easily to observation as such. Think for example about the simple concepts of love, hate, or friendship. In order to solve the "problem" of unobservable instances, we usually *operationalize* them. Instead of observing love as such, we can instead look at cognitive or behavioral operationalizations of love: what do people think or say about the ones they love and how do people behave towards the ones they love? Thus, we can "observe" constructed concepts by observing the manifest operationalizations of the concepts.

Another way to solve the "problem" of non-observable or non-determinable instances in the social world is to employ rhetorics—that is, to determine the power or the reasonableness of arguments as they manifest in language. Since not all instances in the social world fit into the structure of formal logic (deductive reasoning) and since not all instances are readily observable (inductive reasoning) we are dependent on language to solve potential "problems" between people. The field of rhetorics acknowledges this "shortcoming" when it claims that formal-logical arguments are not always employable, whereas rhetorical arguments are more often employable. In rhetorics, formal-logical arguments (deductive reasoning) are known as *syllogisms* (e.g., Smiley 1973), whereas rhetorical arguments are known as *enthymemes* (Jackson and Jacobs 1980). Contrary to a syllogism, which is either true or false—and also "rare" in the real world of 'practical' problems—enthymemes are not true

[10] https://www.britannica.com/science/hypothetico-deductive-method.

or false as such, but are instead more or less convincing as arguments. Since an enthymeme is an argument that does not have the consistency and coherence of strict formal logic, and thus cannot be determined as true or false, it is instead important to generate arguments that are as convincing as possible. Since we are not always able to collect evidence of or verify, all the components (premises) in an argument, we need instead to "make up for" the discrepancy by making the best arguments that we can in the situation.

In the domain of inferences (deductive reasoning and inductive reasoning) this type of *enthymatic* arguments (as they are labeled in rhetorics) are related to *abductive reasoning*—that is, abductive reasoning (Walton 2014) is a "do-your-best" type of inference that is employed when we find ourselves in a situation where we need to handle "syllogisms where one or more premises are missing". In fact, most situations in the real world are like this.

Again, a question arises: How would AI solve formal-logically "wicked" real-life-problems and make up for missing premises in incomplete arguments? Would AI learn how to "make up for" missing premises in arguments and mimic the human way—even if not always successful—of "doing-its-best" by employing the rhetorical devices of abductive reasoning? Would future AI 'transcend' its 'traditional' formal-logic 'way of life' and enter into the domain of human-like reasoning when it infers and subsequently exhibits its behaviour?

5 Conclusion

Even if humans and machines are fundamentally different entities, they are nevertheless similar in one aspect: Both make judgments and inferences that support individual and institutional decisions and both take part in decision making. These decisions have consequences for individual humans as well as for society. It is, therefore, necessary to understand not only human inferences and human behaviour but also Artificial Inferences and Artificial Intelligent Behaviour, so that we can verify the reliability, validity and accountability of the judgments and the decisions that artificial intelligent systems make. In this context, it is necessary to mobilize a comprehensive, unified and multidisciplinary research agenda that includes the breadth of academic disciplines such as natural sciences and technology, humanities and liberal arts, and social sciences. One should not be content with developing artificial intelligent systems, focusing merely on the functionality of the systems. Neither should theoretical analysis be carried out detached from those who develop the intelligent systems. Instead, these two academic disciplines should collaborate and they should let themselves be inspired by the methods used in the cognitive sciences and the behavioural sciences to start studying artificial inference and artificial intelligent behaviour. Systematic empirical studies of artificial intelligent inference and behaviour are necessary to identify and obtain knowledge of AI's *actual* inference and behaviour (AI's judgments and decisions), as a means of obtaining accountability—or at least to identify the current state of accountability—in corporate and government decision making. Moreover, empirical studies of AI's ability to adhere to the specific context it operates within would be of great importance; formally correct reasoning (and judgment) in one context (e.g., screening for diseases) would not necessarily be acceptable in a different context (e.g., rejecting job-applicants) Without such a unified approach, McLuhan's and Postman's analyses of media and technology as autonomous extensions of man, could soon be applied on artificial intelligence, as an equivalent autonomous extension of man and society—beyond oversight, control, accountability, and audit.

The agenda is, however, open as we need first to understand how proven methods of studying human judgment, behaviour and inference can be adapted and used on AI. This has not been done before (only formal methods from computer science are currently being tried out) and thus it represents a novel approach to unravel AI-opacity in the endeavour to provide transparency and accountability. In this joint enterprise, the key to disclosing the link between AI-inference and AI-behaviour is to employ the philosophical *foundations of reasoning* on the study of *AI-behaviour* in specific *contexts* that have specific *purposes* and lead to specific *outcomes*.

Acknowledgements We are very grateful to the anonymous reviewers for their valuable comments that have improved this paper.

References

Ancona D, Bono V, Bravetti M, Campos J, Castagna G, Deniélou PM et al (2016) Behavioral types in programming languages. Found Trends Program Lang 3(2–3):95–230

Anderson FH (1948) The philosophy of Francis Bacon. University of Chicago Press, Chicago

Andrews GR, Schneider FB (1983) Concepts and notations for concurrent programming. ACM Comput Surv (CSUR) 15(1):3–43

Bäck T, Fogel DB, Michalewicz Z (1997) Handbook of evolutionary computation. CRC Press, Boca Raton

Bacon F, Montagu B (1857) The works of Francis Bacon, vol 1. Parry & McMillan, Philadelphia

Baeza-Yates R (2018) Bias on the web. Commun ACM 61(6):54–61

Ben-Ari M (2006) Principles of concurrent and distributed programming. Pearson Education, London

Bergstra JA, Ponse A, Smolka SA (eds) (2001) Handbook of process algebra. Elsevier, Amsterdam

Biamonte J, Wittek P, Pancotti N, Rebentrost P, Wiebe N, Lloyd S (2017) Quantum machine learning. Nature 549(7671):195

Binns R (2018) Algorithmic accountability and public reason. Philos Technol 31:543–556

Biran O, Cotton C (2017) Explanation and justification in machine learning: a survey. In: IJCAI-17 workshop on explainable AI (XAI), vol 8
Boden MA (2016) AI: its nature and future. Oxford University Press, Oxford
Bostrom N (2014) Superintelligence: paths, dangers, strategies. Oxford University Press, Oxford
Bottou L (2014) From machine learning to machine reasoning. Mach Learn 94(2):133–149
Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, Dafoe A, Scharre P, Zeitzoff T, Filar B, Anderson H (2018) The malicious use of artificial intelligence: forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228
Burling R (2005) The talking ape: how language evolved, vol 5. Oxford University Press, New York
Coello CAC, Lamont GB, Van Veldhuizen DA (2007) Evolutionary algorithms for solving multi-objective problems, 2nd edn. Springer, Berlin
Coffey P (1914) Ontology, or, the theory of being: an introduction to general metaphysics, vol 25. Longmans, London
Coffey P (1917) Epistemology; or, the theory of knowledge: an introduction to general metaphysics, vol 1. Longmans, London
Crabtree A, Urquhart L, Chen J (2019) Right to an explanation considered harmful. Edinburgh School of Law Research Paper Forthcoming. Available at SSRN: https://ssrn.com/abstract=3384790 or http://dx.doi.org/10.2139/ssrn.3384790
Dahl ES (2018) Appraising black-boxed technology: the positive prospects. Philos Technol 31:571–591
Danaher J (2016) The threat of algocracy: reality, resistance and accommodation. Philos Technol 29(3):245–268
De Laat PB (2018) Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? Philos Technol 31:525–541
Demuth HB, Beale MH, De Jess O, Hagan MT (2014) Neural network design. Martin Hagan, Boston
Dennett D (1995) Darwin's dangerous idea: evolution and the meanings of life. Simon & Schuster, New York
Dijkstra E (1965) Solution of a problem in concurrent programming control. Commun ACM 8(9):569
Dix A (1992) Human issues in the use of pattern recognition techniques. In: Beale R, Finlay J (eds) Neural networks and pattern recognition in human computer interaction. Ellis Horwood, Chichester, pp 429–451
Dix A (2016) Human-like computing. (Personal report on the EPRC workshop of the same name) http://alandix.com/blog/2016/02/23/human-like-computing/
Dix A (2018) Sufficient Reason. Keynote at HCD for Intelligent Environments, BHCI, Belfast, 3rd July 2018. http://alandix.com/academic/talks/sufficient-reason-2018/
Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608
Dwork C (2011) A firm foundation for private data analysis. Commun ACM 54(1):86–95
Dylan H, Goodman MS, Jackson P, Jansen PT, Maiolo J, Pedersen T (2017) The way of the Norse Ravens: merging profession and academe in Norwegian national intelligence higher education. Intell National Secur 32(7):944–960
Ferrucci D (2011) How it all began and what's next. IBM Research. https://www.ibm.com/blogs/research/2011/12/dave-ferrucci-at-computer-history-museum-how-it-all-began-and-whats-next/
Ferrucci D, Levas A, Bagchi S, Gondek D, Mueller ET (2013) Watson: beyond jeopardy! Artif Intell 199:93–105
Frégnac Y (2017) Big data and the industrialization of neuroscience: a safe roadmap for understanding the brain? Science 358(6362):470–477
Frensch PA, Runger D (2003) Implicit learning. Curr Dir Psychol Sci 12:13–18
Gärling T, Ettema D, Friman M (eds) (2014) Handbook of sustainable travel. Springer, Berlin
Gibson JJ (1979) The Ecological Approach to visual perception. Houghton-Mifflin, Boston
Gilovich T, Griffin D, Kahneman D (2002) Heuristics and biases: the psychology of intuitive judgment. Cambridge University Press, Cambridge
Gordon AD, Henzinger TA, Nori AV, Rajamani SK (2014) Probabilistic programming. In: Proceedings of the 36th international conference on software engineering (ICSE)—future of software engineering track. ACM, pp 167–181
Hennessy M (1988) Algebraic theory of processes. MIT press, Cambridge
Hennessy M, Rathke J (2004) Typed behavioural equivalences for processes in the presence of subtyping. Math Struct Comput Sci 14(5):651–684
Holland JH (1992) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press, Cambridge
Hurlburt RT, Knapp TJ (2006) Münsterberg in 1898, not Allport in 1937, introduced the terms 'idiographic' and 'nomothetic' to American psychology. Theory Psychol 16(2):287–293
Jackson S, Jacobs S (1980) Structure of conversational argument: pragmatic bases for the enthymeme. Q J Speech 66(3):251–265
Jordan MI, Mitchell TM (2015) Machine learning: trends, perspectives, and prospects. Science 349(6245):255–260
Kahneman D (2003) A perspective on judgment and choice: mapping bounded rationality. Am Psychol 58:697–720
Kahneman D, Thaler RH (2006) Anomalies: utility maximization and experienced utility. J Econ Perspect 20(1):221–234
Kahneman D, Tversky A (1972) Subjective probability: a judgment of representativeness. Cogn Psychol 3(3):430–454
Katoen JP, Gretz F, Jansen N, Kaminski BL, Olmedo F (2015) Understanding probabilistic programs. Correct system design—symposium in honor of Ernst-Rüdiger Olderog (vol 9360 of lecture notes in computer science. Springer, Cham, pp 15–32
Kirkham RL, Kirkham RL (1992) Theories of truth: a critical introduction (No. s 401). MIT press, Cambridge
Kneale W (1945). Truths of logic. In: Proceedings of the aristotelian society, vol 46. Aristotelian Society, Wiley, London, pp 207–234
Knuth DE (1973) The art of computer programming vol. 1: fundamental algorithms, 2nd edn. Addison-Wesley Publishing, Boston
Kuhn TS (1962) The structure of scientific revolutions. University of Chicago Press, Chicago
Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of Google Flu: traps in big data analysis. Science 343(6176):1203–1205
LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436
McCarthy J (1987) Generality in artificial intelligence. Commun ACM 30(12):1030–1035
McCarthy J, Hayes PJ (1969) Some philosophical problems from the standpoint of artificial intelligence. In: Webber BL, Nilsson NJ (eds) Readings in artificial intelligence (1981). Morgan Kaufmann Publishers, Los Altos, California, pp 431–450
McLuhan M (1964) Understanding media: the extensions of man. MIT Press, USA
Miller T, Howe P, Sonenberg L (2017) Explainable AI: beware of inmates running the asylum or: how I learnt to stop worrying and love the social and behavioural sciences. arXiv preprint arXiv:1712.00547
Minsky ML (1967) Computation: finite and infinite machines. Prentice-Hall, Englewood Cliffs

Mitchell M (1998) An introduction to genetic algorithms. MIT press, Cambridge
Monroe D (2018) AI explain yourself. Commun ACM 61(11):11–13. https://doi.org/10.1145/3276742
Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592
Newell A, Simon HA (1972) Human problem solving, vol 104, No. 9. Prentice-Hall, Englewood Cliffs
Newell A, Simon HA (1976) Computer science as empirical inquiry: symbols and search. Commun ACM 19(3):113–126
Newell A, Shaw JC, Simon HA (1958) Elements of a theory of human problem solving. Psychol Rev 65(3):151. https://doi.org/10.1037/h0048495
Osawa E, Kitano H, Asada M, Kuniyoshi Y, Noda I (1996) RoboCup: the robot world cup initiative. In: Proceedings of the second international conference on multi-agent systems (ICMAS), Kyoto, Japan, pp 9–13
Pagallo U (2018) Algo-rhythms and the beat of the legal drum. Philos Technol 31:507–524
Parkes DC, Wellman MP (2015) Economic reasoning and artificial intelligence. Science 349(6245):267–272
Pedersen T, Friman M, Kristensson P (2011) Affective forecasting: predicting and experiencing satisfaction with public transportation. J Appl Soc Psychol 41(8):1926–1946
Pedersen T, Johansen C, Jøsang A (2018) Behavioural computer science: an agenda for combining modelling of human and system behaviours. Hum-centric Comput Inf Sci 8(1):7
Popper KR (1972) Objective knowledge: an evolutionary approach. Oxford University Press, Oxford
Postman N (1992) Technopoly: the surrender of culture to technology. Knopf, New York, p 1992
Rahwan I, Simari GR (2009) Argumentation in artificial intelligence. Springer, Berlin
Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C (2017) Machine learning in materials informatics: recent applications and prospects. npj Comput Mater 3(1):54
Rescher N (1973) The coherence theory of truth. Clarendon Press, Oxford, pp 54–64
Robinson OC (2011) The idiographic/nomothetic dichotomy: tracing historical origins of contemporary confusions. Hist Philos Psychol 13(2):32–39
Russell SJ (1997) Rationality and intelligence. Artif Intell 94(1–2):57–77
Russell SJ, Norvig P (2016) Artificial intelligence: a modern approach. Pearson Education Limited, London
Samek W, Wiegand T, Müller KR (2017) Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296
Simon HA (1979) Rational decision making in business organizations (Nobel Prize lecture). Am Econ Rev 69(4):493–513
Simon HA (1996) The sciences of the artificial. MIT press, Cambridge
Sloman SA (2002) Two systems of reasoning. In: Gilovich T, Griffin D, Kahneman D (eds) Heuristics and biases: the psychology of intuitive judgment. Cambridge University Press, Cambridge
Smiley TJ (1973) What is a syllogism? J Philos Log 2(1):136–154
Sutton RS, Barto AG (2018) Reinforcement learning: an introduction, 2nd edn. MIT press, Cambridge
Tversky A, Kahneman D (1973) Availability: a heuristic for judging frequency and probability. Cogn Psychol 5(2):207–232
Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and Biases. Science 185(4157):1124–1131
USACM: US Public Policy Council of The Association for Computing Machinery (2017) Statement on algorithmic transparency and accountability. Association for Computing Machinery, New York
Vardi M (2018) Machine learning and logic: fast and slow thinking. Invited talk at the summit on machine learning meets formal methods. https://easychair.org/smart-program/FLoC2018/SoMLMFM-2018-07-13.html#talk:76996
Wachter S, Mittelstadt B, Floridi L (2017) Transparent, explainable, and accountable AI for robotics. Sci Robot 2(6):eaan6080
Walton D (2014) Abductive reasoning. University of Alabama Press, Tuscaloosa
Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C (2013) Learning fair representations. In: International conference on machine learning, pp 325–333