Hybrid collective intelligence in a human–AI society

Peeters, Marieke M.M.; van Diggelen, Juriaan; Van Den Bosch, Karel; Bronkhorst, Adelbert ; Neerincx, Mark A.; Schraagen, Jan Maarten; Raaijmakers, Stefan

DOI
10.1007/s00146-020-01005-y

Publication date
2020

Document Version
Final published version

Published in
AI and Society

AI & SOCIETY (2021) 36:217–238
https://doi.org/10.1007/s00146-020-01005-y

OPEN FORUM

Hybrid collective intelligence in a human–AI society

Marieke M. M. Peeters[1,4] · Jurriaan van Diggelen[1] · Karel van den Bosch[1] · Adelbert Bronkhorst[1] · Mark A. Neerincx[1,2] · Jan Maarten Schraagen[1,3] · Stephan Raaijmakers[1,5]

Received: 14 January 2020 / Accepted: 8 June 2020 / Published online: 25 June 2020

Abstract

Within current debates about the future impact of Artificial Intelligence (AI) on human society, roughly three different perspectives can be recognised: (1) the *technology-centric perspective*, claiming that AI will soon outperform humankind in all areas, and that the primary threat for humankind is superintelligence; (2) the *human-centric perspective*, claiming that humans will always remain superior to AI when it comes to social and societal aspects, and that the main threat of AI is that humankind's social nature is overlooked in technological designs; and (3) *the collective intelligence-centric perspective*, claiming that true intelligence lies in the collective of intelligent agents, both human and artificial, and that the main threat for humankind is that technological designs create problems at the collective, systemic level that are hard to oversee and control. The current paper offers the following contributions: (a) a clear description for each of the three perspectives, along with their history and background; (b) an analysis and interpretation of current applications of AI in human society according to each of the three perspectives, thereby disentangling miscommunication in the debate concerning threats of AI; and (c) a new integrated and comprehensive research design framework that addresses all aspects of the above three perspectives, and includes principles that support developers to reflect and anticipate upon potential effects of AI in society.

Keywords Artificial intelligence · Hybrid intelligence · Collective intelligence · Human intelligence · Human–AI collaboration · Human–AI society

1 Introduction

Since Alan Turing's ground-breaking work on Artificial Intelligence (AI) in the 1950s, AI research has led to numerous AI demonstrators, steadily invigorating an increasing confidence in the potential of AI (Bughin et al. 2017; Dorado et al. 2018; Loucks et al. 2019). Since the late 1900s and early 2000s, the first practical AI applications have found their way to the market, providing real business value. Present-day overviews of what AI can do are available, for instance, in Newton-Rex (2017) and Dar (2018). On the other hand, the accomplishments of present-day AI also raise concerns about the potentially detrimental impact of AI technology on society. These concerns vary widely, ranging from the imminent advent of rogue Super Intelligence in the near or far future to the dangers of, for example, biased data, prejudiced models, and privacy endangerment.

Almost every day, the news media report on achievements of AI helping to overcome a great variety of real-world problems. One of the key messages in these reports is that innovations in AI technology are able to perform fast and highly accurate computations that surpass human abilities. Examples include breakthroughs in, for example:

- *Medical diagnostics*, e.g. algorithms that are capable of highly accurate recognition of cancerous tissue (Ali 2019);

✉ Jurriaan van Diggelen
jurriaan.vandiggelen@tno.nl

[1] TNO, Kampweg 55, 3769 DE Soesterberg, The Netherlands

[2] Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands

[3] University of Twente, De Zul 10, 7522 NJ Enschede, The Netherlands

[4] Xomnia, Raamstraat 7-I, 1016 XL Amsterdam, The Netherlands

[5] Leiden University, Van Wijkplaats 4, 2311 BX Leiden, The Netherlands

- *Algorithmic trading,*[1] e.g. algorithms that make decisions faster than humans (Crosman 2017);
- *Autonomous driving*, e.g. algorithms that enable cars to predict a crash down the road so as to preventively break autonomously (Davies 2018); and
- *Military warfare*, e.g. algorithms that autonomously select, identify, and engage enemy targets, such as the IAI Harpy[2] (Simonite 2017; Winter 2017).

AI innovations are assumed to result in considerable societal gain, mostly because they perform tasks that are difficult, dirty, dull, or dangerous for humans. Automating such tasks would either no longer require humans to perform them at all, or enable humans to still perform them, but to do so more effectively or more efficiently (Kunze et al. 2018). To some, the achievements of current AI technology are construed as only the beginning of a fantastic future, whereas others have been eager to point at potential limitations and threats (Hadfield-Menell et al. 2017; Sharkey 2017; van Wynsberghe and Robbins 2018). In this paper, we identify roughly three positions in the current debate about AI. These positions are briefly introduced here, and will be analysed in more depth in the sections to come.

The first view can be attributed to those with high expectations of AI. People who take this position in the debate tend to stress the attainability of omnipotent AI and its profound consequences for humanity as we know it (Bostrom 2016). Within this position, different opinions exist regarding the consequences of potential artificial super intelligence. Some anxiously warn against the dangers of artificial super intelligence, and stress the need to implement safeguards to ensure that future AI systems will remain benevolent and beneficial to humanity. Others are less concerned and believe that AI itself will be able to solve the dangers that face humankind in the next few decades: “If AI can perform its tasks at superhuman levels of performance, why then not assign many or all tasks to AI?” In this paper we refer to this view as the *technology-centric* perspective.

A second view can be roughly attributed to those who foresee a predominantly negative impact of AI. People who take this stance expect insurmountable problems from assigning societal activities to AI. They raise questions like: “What impact would a gradual shift towards automated labour have for humanity’s sense of fulfilment and meaning?”; “What happens if we would gradually delegate responsibility and decision-making to AI; would humans become insignificant and subordinate?”; “If AI would carry out most of the tasks that shape our society, what would happen to our autonomy, or our countries’ sovereignty even?”; and “Would the proliferation of AI always optimise towards societal benefit, or could it also lead (perhaps unknowingly) to detrimental effects that degrade rather than improve our societal values?” Throughout the rest of this paper, this view will be referred to as the *human-centric* perspective.

In addition to the technology-centric and the human-centric perspective, we identify a third position in the debate, which we call the Collective Intelligence perspective. Collective Intelligence originally comes from the idea that humans can connect in a way that allows them to collectively act more intelligently than any individual person (Engel et al. 2014; Henrich 2015; Sloman and Fernbach 2018; Sutton et al. 2010; Theiner et al. 2010; Woolley et al. 2010). Although the term Collective Intelligence originally referred to groups of people, in recent years, the concept has been adopted and gradually extended to refer also to the collective groups of people and intelligent technology (Malone 2018; Malone and Bernstein 2015; Mittrick et al. 2019; Smirnov and Ponomarev 2019). In the literature, alternative terms have been used to describe collective intelligence, for example, “intelligence amplification” (Ashby 1961), “intelligence augmentation” (Engelbart 1962; Sesay and Steffen 2020), “symbiotic intelligence” (Licklider 1960), “extended intelligence” (Clark and Chalmers 1998; Adamson et al. 2019), and “hybrid intelligence” (Dellermann et al. 2019). We prefer, however, the term “Collective Intelligence”. The Collective Intelligence perspective is consistent with the dominant systems-of-systems perspective in engineering, as becomes clear from the following quote:

> Instead of thinking about machine intelligence in terms of humans vs. machines, we should consider the system that integrates humans and machines – not artificial intelligence but extended intelligence. Instead of trying to control or design or even understand systems, it is more important to design systems that participate as responsible, aware, and robust elements of even more complex systems (Ito 2019, p. 1).

The field of artificial intelligence (AI) has, for decades, attempted to create computer programs that can behave as intelligently as humans. Achievements of AI tend to be considered a breakthrough only when they can be accomplished independently, without human involvement (at least at runtime/during operation). Researchers working in the field of collective intelligence, however, state that it should not be considered cheating when people are allowed to help a program while it is running.[3] They argue that solving today’s most critical and difficult real-world challenges needs teams

[1] Algorithmic trading (2019). Retrieved from https://en.wikipedia.org/w/index.php?title=Algorithmic_trading&oldid=883815399.

[2] https://www.iai.co.il/2013/36694-16153-en/Business_Areas_Land.aspx.

[3] https://cci.mit.edu/.

Fig. 1 Tenets of techno-centric view

T1 When sufficiently developed, AI technology can be applied to solve any problem.

T2 AI technology may introduce additional problems which can, in turn, be solved by AI.

T3 As the maturity of AI increases, there will be less need for user interaction.

T4 Current AI technology has only reached a fraction of its full development potential.

T5 AI has vastly more potential than human intelligence.

consisting of human and artificial agents, working together (Malone 2018).

The current paper offers the following contributions: (a) a clear description for each of the three perspectives, along with the history and background; (b) an analysis and interpretation of current applications of AI in human society according to each of the three perspectives, thereby disentangling miscommunication in the debate concerning threats of AI to human society; and (c) a proposal for a new research paradigm and framework to address all aspects of the debate, and the three perspectives. This aims to facilitate development of new research methods to investigate and moderate the potential threats of AI to human society from different angles and at different systemic levels.

The structure of this paper is as follows:

- Section 2 provides an overview of the three perspectives on AI innovations and their implications for society: the human-centric perspective, the technology-centric perspective, and the collective intelligence perspective.
- Section 3 presents an analysis of recent AI innovations in a range of application domains, resulting in arguments for and against each of the three perspectives presented in Sect. 2.
- Section 4 presents a design framework that adopts elements from the three perspectives in a 360° angle view on AI innovations. The framework supports measuring, predicting, and mitigating (unwanted) effects of AI at different levels of society.
- Section 5 presents our concluding remarks.

2 Three perspectives on AI

Recently, public media and scientific literature offer ample opportunities for debate about the potential impact of AI on society. The debates mostly revolve around the question: “How will human intelligence relate to artificial intelligence within the next few decades?”. Obviously, we can encounter as many opinions as there are experts. Nevertheless, we can also begin to observe several lines of thought that are shaping up the debate. Without implying that everyone fits exactly within one of these categories, we propose the following three perspectives on AI:

1. *The technology-centric perspective*, which holds that true intelligence can ultimately be found only in well-developed and matured (general) AI systems. Humans are biologically constrained in their information processing and reasoning capabilities, and display many types of cognitive bias, while computers provide virtually endless opportunities to develop rational intelligence at and beyond the human level.
2. *The human-centric perspective*, which holds that true intelligence can ultimately be found only in human beings and (potentially) other intelligent living creatures. AI can help humans to reach their full potential, but will by nature not be able to develop certain essential qualities found in humans, such as moral reasoning or empathy. Due to this incapability, AI may cause danger to human well-being.
3. *The collective intelligence perspective,* which holds that true intelligence can ultimately be found only in the collective of multiple interacting entities. In isolation, the intelligence of the individual human and AI entities within a system is extremely limited. True intelligence emerges when multiple entities collaborate over longer periods of time.

The next sections describe these three positions and their origins in further depth.

2.1 The technology-centric perspective

The technology-centric view on AI (or `techno-centrism’) is grounded in a belief in the huge and continuously expanding potential of AI, as exemplified by the ability of current AI systems to outperform humans in various tasks (Bostrom 2016). Although followers of techno-centrism admit that new technologies can introduce additional problems, they are also eager to point out that these problems can again be solved by applying additional technology. Whereas the different perspectives on AI are just beginning to take shape, the technology-centric perspective is articulated more explicitly in the environmentalist movement (Bailey and Wilson 2009). In the debate on climate change, followers of techno-centrism are *in favour* of technological solutions, such as building electrical cars and CO_2 capture, and are generally *dismissive* of behavioural solutions, such as discouragement of using high carbon-producing activities like air travel and meat consumption.

Figure 1 presents some of the main tenets that underly the techno-centric perspective (referred to as T1–T5).

Followers of techno-centrism are generally optimistic about the expected advancements in AI. It is thought to be only a matter of time before AI will equal and even surpass human intelligence on many (or all) fronts. This will confront us with the problem of dealing with entities more intelligent than ourselves, who can make decisions and take actions that may be incomprehensible to us (Brynjolfsson and Mitchell 2017). An important advantage attributed to AI is that it does not suffer from the same limitations on information processing as humans, such as limited (working) memory, biases and heuristics, fatigue and stress, and social pressures. As a result, AI is believed to be free from these "human brain"-related errors in decision-making. Furthermore, AI can be pre-programmed to pursue clear mathematically defined goals while considering legal and ethical constraints (Bostrom and Yudkowsky 2014). Oftentimes, AI is described as being perfectly rational, as opposed to humans who suffer from all sorts of biases and cognitive limitations (Russell 2019).

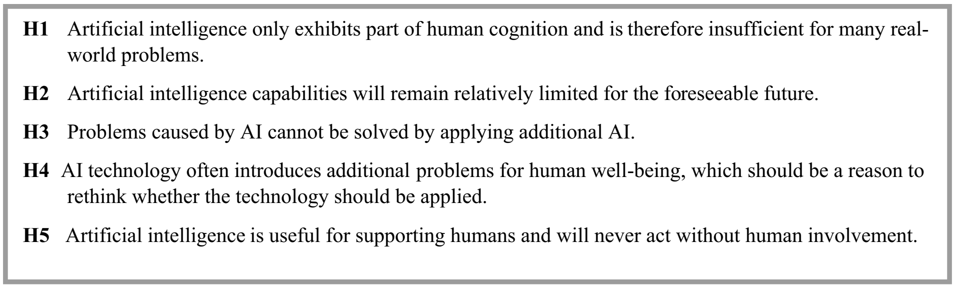
Not only are techno-centrists optimistic about the potential uses of AI; but also techno-centrism often goes together with scepticism towards human abilities to make fair judgments. For example, Kahneman (2011) has demonstrated that human decision-making can be severely flawed, as humans tend to use heuristics that are suboptimal and are likely to produce biased outcomes. Techno-centrists argue that this human deficiency may carry over to the decision-making processes of AI. For example, humans may contribute to *selection bias* when selecting the training data for self-learning AI (Lloyd 2018), *label bias* (Jiang and Nachum 2019) when pre-labelling raw data for AI to learn from, and -on the technical level- may introduce *inductive bias* (Hüllermeier et al. 2013) into an AI system when developing its mechanisms for generalisation over new data. In adversarial machine learning (Papernot et al. 2017), exploiting inductive bias of machine learning algorithms by malevolent humans contributes to undermining AI. Furthermore, intrinsic obscureness of complex AI algorithms (such as deep, temporal neural networks with often millions of parameters) obfuscates the *data auditability* (Raaijmakers et al. 2017) of AI for humans, increasing the risk of black box biased AI. Techno-centrists argue that bias introduced by humans imposes a risk so extensive that it is better to exclude human influence from the AI decision-making process as much as possible (Miller 2018).

Some proponents argue that although current AI applications still have a narrow scope, they will soon evolve into Artificial General Intelligence (AGI), meaning that it can perform any intellectual task that a human can. Once AGI is achieved, Artificial Super Intelligence (ASI) soon becomes within reach (Bostrom 2016; Kurzweil 2005), because the AGI can apply its own intelligence to rewire itself into a system that is even more intelligent. A less far-reaching form of superintelligence (i.e. narrow superintelligence) can be understood as a narrow AI reaching super-human performance within a specific task domain.

Advocates of the technocentric perspective have high expectations of AI, and they envision a declining role for humans in task execution and society in general. The argument is that, if AI performs at a superhuman level, human involvement in decision-making can only worsen or slow down performance. At some point, humans will become incapable of being involved as they can no longer understand the computer's super intelligent line of reasoning. Therefore, humans should preferably be kept out of the loop, and a technological solution should instead be developed to ensure that the AI does not act against humanity's interests (for example by an ethical utility function (Bostrom 2016)).

Techno-centrists assume that Artificial Super Intelligence (whether narrow or general) will have a huge impact on humanity, although there is no consensus on what the outcomes may be. Predictions range from, on the positive side, more humane robotic warfare, safer transport, the possibility of space colonization, and on the pessimistic side, to mass unemployment, and even human extinction. The possibility of AI causing human extinction has even incited a new philosophical movement, namely the transhumanist movement (Kurzweil 2005). This movement holds that technology may be used to transform humans into an upgraded species, and that this should not necessarily be a bad thing.

Whereas some of these visions on the future of AI might strike the reader as science fiction, they are a substantial part of the current debate on where AI technology is heading. Discussions on AGI and superintelligence are nothing new (e.g. Searle 1980), and forecasts on AI developments have a long history of being overly optimistic about (soon-)to-be-achieved capabilities. Nevertheless, the debate has been revitalised since distinguished figures, such as physicist Stephen Hawking and business magnate Elon Musk, signed an open letter (Future of Life Institute 2015) in which they warned (among other things) against the risks of artificial superintelligence. Even so, a recent study among twenty-three of the world's foremost AI researchers and entrepreneurs showed that opinions on when Artificial General Intelligence (AGI) might be available are highly divergent. Some think it may be achieved in our lifetime, others think it will not (Ford 2018). Thus, there is no consensus among AI researchers on when AGI will be reached, if at all.



H1 Artificial intelligence only exhibits part of human cognition and is therefore insufficient for many real-world problems.

H2 Artificial intelligence capabilities will remain relatively limited for the foreseeable future.

H3 Problems caused by AI cannot be solved by applying additional AI.

H4 AI technology often introduces additional problems for human well-being, which should be a reason to rethink whether the technology should be applied.

H5 Artificial intelligence is useful for supporting humans and will never act without human involvement.

Fig. 2 Tenets of human-centric perspective

2.2 The human-centric perspective

The human-centric perspective on AI (or 'human-centrism') views AI primarily as a tool for improving the performance, safety, and well-being of humans (Baum 2017; Russell et al. 2015), but not one that will eventually replace humans. According to this view, AI may be used for tasks and services that humans are not willing or able to perform. For example, dirty, dull, or dangerous tasks, handling of large volumes or high velocity of data, or supporting people that require help or care (Brynjolfsson and McAfee 2014). However, the human-centric view also stresses the limitations of AI (Brynjolfsson and Mitchell 2017; Ng 2016): AI is mostly regarded as a technology with a restricted capability envelope (Endsley 2018), that suffers from errors (Yampolskiy and Spellchecker 2016), and that is inherently sensitive to biases in the input data (Osoba and Welser 2017). Human-centrists also argue that AI cannot reason as humans do, nor do they have the same knowledge available for making judgments (Legg and Hutter 2007). Proponents of human-centrism believe that AI should, therefore, be applied only after serious consideration of all its potential benefits, drawbacks, and disadvantages. Although the human-centred view on AI is diverse, we can extract some commonalities that are relevant to consider when proposing an AI engineering method, as provided in Fig. 2.

Human-centrists are convinced that human intelligence and artificial intelligence are different by nature, and therefore cannot substitute one another. The origin of this idea can be led back to Fitt's list in the 1950s which provides guidelines for function allocation based on what *men are better at*, and what *machines are better at* (Fitts 1951). Despite having received extensive criticism over the decades, the idea that some functions can better be performed by humans remains popular (de Winter and Dodou 2014). Task typologies and taxonomies are commonly linked to required capabilities to decide whether to assign a given task to humans or to machines. There is general consensus, at least among human-centrists, that current AI capabilities are specialist and domain specific in nature, causing their applicability to be restricted to highly circumscribed task domains or even situations, and limiting their adaptivity to the degrees of freedom accounted for within the given application (Schank 2017). Following this line of reasoning, AI systems function well in environments in which they are trained, yet become brittle in novel situations. For example, real-world environments tend to be 'messy', containing factors of influence that are ill-defined, inherently uncertain, or difficult to foresee (e.g. Woods 2016). This argument has important ramifications for the use of AI in, e.g. self-driving cars (Surden and Williams 2016) and military applications (Department of Defense 2015). Although human expertise is also domain-specific to a large extent (Feltovich et al. 2006), humans are better capable of adapting to novel domains than AI systems can (Klein et al. 2020). Human-centrists, therefore, conclude that humans should remain in control of decision-making and task execution to compensate for AI's narrow specialism, and its associated rigidity and brittleness.

Also, with respect to morality, human intelligent capabilities are considered indispensable for decent ethical deliberation and decision-making. Because ethical deliberation and decision-making is paramount to human existence and well-being, application of fully autonomous AI in ethically sensitive domains is unacceptable. For example, peace organisation PAX states about autonomous weapon systems: "A machine should never be allowed to make the decision over life and death. This goes against the principles of human dignity and the right to life. This decision cannot be reduced to an algorithm. Outsourcing this decision would mean outsourcing morality."[4]

Human-centred researchers refute techno-centrism, especially the claim that AI outperforms humans as AI is believed to be free of human bias and capable of perfect reasoning. Human-centrists argue that heuristics, in their original meaning of 'rules-of-thumb', were never meant to be optimal (e.g. Polya 1945; Sloman and Fernbach 2018). A second argument used is that biases as described by, for instance, Kahneman, are primarily artefacts of controlled laboratory research carried out with naïve participants: Studies that have tried to replicate this research with problem

[4] https://www.paxforpeace.nl/media/files/pax-ten-reasons-to-ban-killer-robots.pdf.

Fig. 3 Tenets of collective intelligence perspective

C1 Intelligence should not be studied at the level of individual humans or AI-machines, but at the group level of humans and AI-machines working together.

C2 Increasing the intelligence of a system should be achieved by increasing the quality of the interaction between its constituents rather than the intelligence of the constituents themselves.

C3 Both human as well as artificial intelligence can be regarded as very shallow when considered in isolation.

C4 No AI is an island.

statements couched in familiar terms, have invariably found that the biases disappeared altogether or were much less pronounced. For example, in one study participants were instructed to verify that a set of cards was consistent with a rule stating that people are not allowed to drink alcohol under a certain age; one side of the card stated the drink a person was having, and the other side of the card revealed that person's age. This task was a more familiar variation on a similar lab experiment instructing people to verify that a set of cards was consistent with a rule stating that a vowel must be assigned only to cards with an odd number on the back. People performed much better in the real-world version (the age restriction on alcohol) than they did in the abstracted version of the experiment. The third argument brought forward by human-centrism is that biases are frequently measured against a normative yardstick that is inappropriate for people, such as formal logic or Bayesian statistics (Klein et al. 2020). Following this line of reasoning, both heuristics and biases are better viewed as effective adaptations of humans to reason about and act within their natural environments. Ironically, while it is clear that AI can theoretically be programmed to apply normative 'bias-free' reasoning to problems, the deep neural nets that represent the current state of the art are trained with, and therefore completely depend upon, large (often handcrafted) labelled datasets that will, as a rule, be biased. As a result, it has been shown that such systems may discriminate individuals on the basis of race, gender, or sexual orientation, thus reproducing the same prejudices as the humans who originally produced the data on which the algorithms were trained (O'Neil 2017).

In society, the human-centric perspective often manifests itself in reaction to the technological disruptions that influence or manipulate humans on a large scale (e.g. social media, smartphones, and AI algorithms). For example, Rushkoff puts forward that we live in a world dominated by data gathering and algorithmic optimization, and he pledges to "join team human" (Rushkoff 2019), by relying on human values such as creativity, social connections, and respect. Other advocates of the human-centric perspective refrain from technologies such as social media, smartphone usage, or search engines and other technologies that do not value privacy. In her book "Weapons of math destruction", O'Neil (2017) warns against the rise of oversimplified (data-driven) models that are being used for loan assessment and recruitment. She considers the effects of such models on society to be devastating, because these models affect large proportions of the population, and are non-transparent in their decision-making.

The concerns of the human-centred perspective are prominently expressed in the European roadmap for human-centred AI.[5] The first line of the mission statement reads: "CLAIRE will focus on trustworthy AI that augments human intelligence rather than replacing it, and that thus benefits the people of Europe."

2.3 Collective intelligence perspective

Proponents of the collective intelligence perspective stress that humans and AI can connect in ways that allow them to collectively act more intelligently than any of the individual entities alone. Collective Intelligence (CI) can be defined as "shared or group intelligence that emerges from the collaboration, collective efforts, and competition of many individuals."[6] Originally, CI researchers primarily aimed to study how groups of people act and think "as a whole", e.g. using various coordination and decision-making mechanisms. The field dates back to 1907, when statistician Francis Galton conducted his famous experiment asking a large group of participants to estimate the weight of a cow (Wallis 2014). The results showed that, although none of the participants provided the exact right answer, the average of estimations was less than 1% away from the actual value. Even though the field originally focused on groups of people, in recent years, the field has gradually expanded to also include artificially intelligent systems as group members. Researchers investigating this Hybrid Collective Intelligence "explore how people and computers can be connected so that—collectively—they act more intelligently than any person, group, or computer has ever done before" (Malone 2018). Statements that can be regarded as typical for the Collective Intelligence approach are listed in Fig. 3.

[5] https://claire-ai.org/.

[6] https://en.wikipedia.org/wiki/Collective_intelligence.

Even though the results obtained by Francis Galton's experiment are frequently cited as showing the potential of CI, it is also clear that most efforts to make individuals think collaboratively as a group are much more challenging. The CI perspective has often been applied to better understand why some organisations are more effective than others, or to better understand the cause of an accident, such as, for example, the accident with the Columbia Space Shuttle (Surowiecki 2005). Common factors that determine the level of CI are, for example, the level of interconnectedness, diversity, hierarchy, and critical culture. On a societal level, CI could be applied to design a democracy in such a way that it expands the brainpower of a society instead of dumping it down (Mulgan 2017).

Advances in internet technology have renewed interest in collective intelligence yielding novel applications such as crowdsourcing to build software, encyclopaedias (e.g. Wikipedia), and digital maps. Important design considerations to make such systems work well are, for example, incentive mechanisms (for individual contributors, but also business models for companies), fault correction mechanisms, and sabotage prevention (Awad et al. 2020). Note that this underlying technology is, in itself, not AI. Rather, it should be viewed as infrastructure that results in more intelligence on the collective level.

When a group that exhibits intelligent behaviour consists of humans and AI systems, we can speak of collective hybrid intelligence (Kamar 2016). This is also known as a joint cognitive system (Hollnagel and Woods 2005), or a human agent team (HAT). Researchers studying human–agent teaming argue that the combination of AI, humans, and social artificial intelligence (van Diggelen et al. 2018) is needed to obtain a truly intelligent system. In such a system, humans can compensate for a machine's weaknesses and vice versa. Although AI may function more or less autonomously, a tenet of the CI perspective is that all AI systems must at some point interact with humans. Therefore, "no AI is an island" (Johnson and Vera 2019).

Collective intelligence can be identified at multiple levels in a system:

- at the *dyadic level*, e.g. a human doctor and a decision support system trying to decide upon a diagnosis and the best course of action;
- at the *team level*, e.g. a swarm of drones, various human operators, and a team leader offering protection for a village under attack by hostile forces;
- at the *organisational multi-team level*, e.g. multiple Urban Search and Rescue teams operating at various locations in a hazard area, and taking instructions from a central control unit overseeing the mission as a whole and handing out strategic orders to each of the teams; or
- at the *societal and cultural level*, e.g. multiple systems and infrastructures interacting with one another, together resulting in emergent effects stretching beyond the boundaries of the organisation itself and into the real world. Examples are disruption of traffic infrastructure, discrimination against groups of people, and/or hazardous effects on climate change and other environmental aspects.

The collective intelligence perspective has proven useful not only to identify opportunities, but also to identify problems and even threats to human well-being. In the following, we present several examples of such problems and threats.

A first example comes from a study by Van Panhuis et al. (2014). They conducted a systematic review regarding barriers to data sharing in public health. When looking at the collective level, it is plain to see that data sharing is beneficial for the system as a whole, as it allows for faster, better, and more inclusive ways of developing and combining knowledge regarding health issues and potential solutions to health threats. However, at the organisational and individual level there are various reasons not to share data, like: the risk of data being used to name and shame institutions that are lagging behind on health policies or programs; or the risk that shared data are used by a (bigger) competitor to reap the benefits before the original collector of the data is able to do so. At a societal level, a barrier to share health data may be, for example, the fear of economic damage due to a drop of tourism and trade in case of an epidemic or pandemic. There may also be political barriers, such as a lack of trust in the people receiving the data, or a lack of guidelines on sharing data; legal barriers, such as copyright or privacy laws causing individuals to be cautious regarding data sharing; or even ethical barriers, such as fear of disproportionality (e.g. the benefits of data sharing are not proportional to the risks regarding privacy or security) or lack of reciprocity (e.g. sharing data with the other party, while the other party does not share their data in return). This example on sharing of public health data shows that an analysis at the collective system allows for the identification of structural problems and perverted incentives. It convincingly illustrates how different interests may ultimately lead to behaviour that is disadvantageous at the collective level, i.e. leading individual healthcare professionals and health organisations to refrain from sharing their data to improve public health.

The second example comes from the book "The knowledge illusion: why we never think alone" (Sloman and Fernbach 2018). The authors argue that human achievements are mostly the result of collective intelligence, each person continuing the work of their predecessors and learning from interactions and discussions with their peers. Sloman and Fernbach make the case that individual humans have a very shallow understanding of most things (viz. tenet C3), and may have a deep understanding of some things within their field of expertise. Yet, for the larger part, intelligence resides in the collective mind. Modern information technology has led to an immense increase in connectedness and thereby caused this phenomenon to be even more prevalent: the internet offers a huge external storage of knowledge, facts, ideas, and theories that people use on a daily basis. *The knowledge illusion* refers to the phenomenon that most people are unaware of the extent to which they rely upon collective intelligence and, as a result, people's tendency to overestimate their individual knowledge and understanding. This "fallacy" has a dramatic impact on the way people design, develop, and use AI systems, especially when looking at this from a collective intelligence perspective. People often mistake the solutions provided by intelligent systems for thoughts developed by themselves. The boundaries between the products of thinking and the outcomes of artificial algorithms become diffuse. This tendency of people may impel precautionary measures in human–AI cooperation. For example, Kamphorst and Kalis (2015) have argued that designers should be mindful of the risks when offering users of autonomous e-coaching systems a set of options, "especially those that combine persuasive techniques such as reduction, tunneling, tailoring and self-monitoring with personalization to actively influence their user's behavior in order to achieve lasting behavior change (p. 77)". This example shows how well-intended technology, such as e-coaching, runs the risk of becoming an instrument of mass manipulation, a risk that evolves due to tendency of humans to overestimate their own intellect. Such potential effects of intelligent technology become especially visible when looking at it from a collective intelligence perspective.

The third example describes how intelligent technology can result in one group of people controlling and manipulating another group of people. In platforms such as Uber and Deliveroo, humans are faced with the effects of AI systems while being unable to exert control over it. For example, the AI determines fares and rides, while the drivers have limited to no control. Douglas Rushkoff predicted the problem of loss of human control over technology almost a decade ago in his thought-provoking book "Program or be programmed: Ten commands for a digital age" (Rushkoff 2010). In his book, he states:

> Our enthusiasm for digital technology about which we have little understanding and over which we have little control leads us not toward greater agency, but toward less… We have surrendered the unfolding of a new technological age to a small elite who have seized the capability on offer. (p. 140)

The problem observed here is that some AI systems (such as platform work systems) are designed to affect or even manipulate a large group of people, but at the same time leave very little possibilities for that same group of people to influence the behaviour of that AI system. It goes without say that such a mechanism may harm people's autonomy and as a result should be carefully regulated (Kamphorst 2012). Once more, when analysing technology from a collective intelligence perspective, such effects are more likely to become clear. These undesired outcomes are not a fault of the technology itself, nor do they necessarily imply that all humans involved in the system need to be able to exert more control. But the observation that at the collective level, the system fails to establish fairness and autonomy for those involved, should be taken as a warning to make changes in the design and implementation.

The insights obtained from examples such as outlined in the above can have huge implications. Analysis from a collective intelligence perspective supports developers to design systems in which decisions are made and interpreted as intended, supports the anticipation, detection and resolution of potential misconducts, and supports a proper implementation of responsibility and accountability in the organisation and in society as a whole. To design and develop complex collective intelligence systems that allow control at all involved levels, there is a need for validated patterns for interaction, teaming, coordination, and decision-making (van Diggelen et al. 2018, 2019).

3 AI manifestations in current and future society

3.1 Examples of AI applications in society

Modern society has many examples of AI applications. They differ, among other things, in their maturity, ease of use, purpose, and added societal value. In the following subsections, we present some of the major AI developments in (1) games (2) intelligent conversational agents and personal assistants, (3) (semi-)autonomous cars, (4) art and social media, (5) stock trading agents and fintech, (6) logistics and decision support, and (7) military systems and robotics. Within each subsection, we analyse the relevant developments by presenting a brief overview of AI applications in the respective domain area. In Sect. 3.2, after presenting AI developments in all domains, we behold the entire body of AI developments, and discuss them in the light of the three perspectives to provide insights into the future of AI and its (potential) impact on society.

3.1.1 Artificial intelligence in games

One of the early achievements made by AI has been in competitive board games. One could argue that it all started with Deep Blue 2 beating the then-world champion Garri

Kasparov (Campbell et al. 2002). From there on, the world witnessed a series of ever more impressive accomplishments of AI in mastering games. In 2011 IBM Watson won Jeopardy (Chen et al. 2016; Ferrucci 2012; High 2012). In 2014, DeepMind[7] developed AI able to play a variety of seven different arcade games, such as Pong. Most recently DeepMind trained one of its systems to play Quake III Arena and currently the DeepMind team is working on their AI to play StarCraft II, a real-time strategy computer game, and Hanabi, a collaborative card game that relies on each player's ability to reason about other players' reasoning given each player's potential information state (Hao 2019). In 2016, AlphaGo beat one of the highest-ranked players in the world at the game of Go, and in 2017 AlphaGoZero beat the original AlphaGo system (Silver et al. 2017). By now, AlphaZero has taught itself to play chess and shogi as well. Another great accomplishment was that of Libratus in 2017, which won a poker tournament playing against four top-class human poker players (Brown and Sandholm 2018). Recent advances in the field of game-playing AI are often used as supportive evidence of the techno-centric view on AI. In the years after computer Deep Blue beat human chess champion Gary Kasparov, a hybrid system consisting of a human supported by a computer was still capable of beating the best solo chess computer (Case 2018). However, due to great technological advancements, when it comes to playing chess, computers now vastly outperform any human or human–AI team. Alpha Go Zero taught itself to learn the game of Go by playing against itself. At one point, the computer famously made a move that no human Go player would ever play, but which turned out to be brilliant (Metz 2016). In fact, training AlphaGo on human sample games turned out to accelerate the learning rate of the system, but led to decreased performance. These examples are put forward as evidence for the technology-centric perspective that eventually technological progress will make human thinking obsolete.

3.1.2 Intelligent personal conversational assistants

Another area in which Artificial Intelligence has made great progress is natural language processing and synthesis, enabling innovations like intelligent personal conversational assistants. Ever since the first conversational AI, Eliza (Weizenbaum 1966), the promise of AI providing assistance through voice has long been considered a more natural and intuitive way of interaction. In comparison to Eliza, modern conversational interfaces show remarkable performance. Nowadays, producers of consumer goods and web services can choose to line up with platforms like Google Assistant,[8] Microsoft Cortana,[9] Apple Siri,[10] or Amazon Alexa,[11] allowing their customers to instruct appliances through voice commands and receive information in the form of voice messages, resulting in brief dialogue flows. Some of these personal conversational assistants display highly natural emotions in tone of voice, others provide multi-modal information in answer to a question, and most of them support a wide variety of services, such as online shopping, setting timers, writing emails, or telling jokes (López et al. 2018). Yet on the other hand, some assistants break down when asked to perform tasks not supported (yet) by their manufacturers, or lose naturalness in response or tone of voice (López et al. 2018). The technological advancement in the area of intelligent personal conversational assistants has also led to more controversial applications, such as: Hello Barbie by ToyTalk[12] (Holloway and Green 2016), which ignited debates about privacy and child rearing; and Twitterbot Tay, created by Microsoft and taken down after one day as Twitter followers successfully tested Tay's limits by "feeding" it with racist, misogynist, and antisemitic slurs, causing the chatbot to utter increasingly violent and hateful expressions on Twitter (Horton 2016; Price 2016). Other examples illustrative of the challenges related to smart personal assistants have been Amazon Alexa recording a private conversation and sending it to a random contact (Wolfson 2018), and the commercial created by Burger King,[13] exploiting Google Now's activation using the words "OK, Google" causing the personal assistant to read out loud the Wikipedia page for the Whopper (Maheshwari 2017).

3.1.3 (Semi-)autonomous cars

Currently, high-end cars from mainstream car manufacturers offer level 2 autonomy on the road, i.e. "hands-off", meaning that the car will take full control of accelerating, braking, and steering, while the driver monitors the driving and remains prepared to intervene when needed. Examples of this are Tesla's model X and model 3[14] and the Volvo XC60.[15] Gradually, upcoming models, such as the Audi A8 (Basem 2018) and Waymo's full autonomous taxi service

[7] https://deepmind.com/.

[8] https://assistant.google.com/.

[9] https://www.microsoft.com/en-us/cortana.

[10] https://www.apple.com/siri/.

[11] https://www.amazon.com/Amazon-Echo-And-Alexa-Devices/b?ie=UTF8&node=9818047011.

[12] https://www.toytalk.com/product/hello-barbie/.

[13] https://www.youtube.com/watch?v=zedWOAtLdn4.

[14] https://www.tesla.com.

[15] https://www.volvocars.com/intl/cars/new-models/xc60/specifications/features.

cars (Hawkins 2018; Sage 2018), also include the potential of level 3 autonomy, i.e. "eyes-off", allowing drivers to look away (e.g. nap, read, or watch a movie) and be alerted by the car in time to take back control from the car's AI (also see Davies 2018). As of now, car companies like General Motors (LeBeau 2018), Waymo,[16] Nissan,[17] and Ford,[18] forecast that level 4 autonomous driving is to be expected within 5 years from now; in that case, cars will be able to drive autonomously in predetermined areas and under certain conditions (not just anywhere and anytime). As of late, debates are heating up about the behaviour of autonomous cars. Important questions raised in this context boil down to a version of the trolley problem: If a collision is inevitable, and a car has the time and resources to compute within fractions of a second the potential damage it will do when hitting one object or another, what object should it choose and based on what calculation (Lin 2013, 2014a, b; Maurer et al. 2016)? The big challenge in this dilemma is that the behaviour of the car in such a situation depends on the choices made by its programmer(s), either deliberately or unknowingly. This raises question, such as whether the programmer is eligible to make such a decision, and if not the programmer, then who?, whether it would be wise to have all cars behave in the same manner given a certain collision scenario, potentially causing a specific group of people to be "targeted" in all those cases, and whether cars should be making such decision "by themselves" in the first place.

3.1.4 Creative content and (social) media

Another important area where AI has shown tremendous progress is (social) media. Well-known examples include Google Personalised Search,[19] Facebook News Feed (Constine 2016), and Twitter Timeline,[20] providing their members with personalised information streams mixing family photos, friends' status updates, advertisements, and magazine and newspaper articles. Spotify[21] and Netflix[22] are more targeted platforms offering their customers personalised music playlists and recommendations for video content, respectively. Not only is AI used to organise and select existing information to present to the reader, AI is also used to create the content itself. For example, RADAR[23] is a tool used to automatically create news articles; IBM Watson Beat,[24] Google nSynth Super (Deahl 2018), MeloDrive,[25] and JukeDeck[26] support musicians in creating symphonies and songs; researchers are working on AI that is capable of automatically writing novels (Streitfeld 2018), and Textio[27] aids recruiters in writing texts for job vacancies. Recently, personalised news feeds have gathered negative publicity, as they were presumably used to manipulate people's political opinions (Pariser 2011; Isaak and Hanna 2018; González 2017) and shopping and buying behaviour (Rushkoff 2010). People's behaviour is increasingly affected by algorithms that select and present news articles confirming their belief systems, unintentionally creating so-called filter bubbles and self-fulfilling prophecies through feedback loops, which in turn can lead to political polarization (Rushkoff 2010). Especially disconcerting is a new phenomenon called "deep fakes" using generative adversarial networks, an AI technique, allowing one to combine and superimpose existing images and video onto source images or video materials (Metz 2018). Especially, the combination of automated information generation, selection, and presentation is the cause for many to sound the alarm bell on a potentially massive and powerful propaganda and mass-manipulation machine (Woolley and Howard 2017; Morgan 2018). As Lanier (2018) puts it in an excerpt from his latest book: "Algorithms gorge on data about you, every second. (…) All these measurements and many others have been matched up with similar readings about the lives of multitudes of other people through massive spying. Algorithms correlate what you do with what almost everyone else has done. (…) So-called advertisers can seize the moment when you are perfectly primed and then influence you with messages that have worked on other people who share traits and situations with you. (…) What might once have been called advertising must now be understood as continuous behaviour modification on a titanic scale."

3.1.5 Stock trading agents and FinTech

In the financial sector, AI has been around for quite some time, where it has been used for stock trading, such as Kavout,[28] Green Key,[29] or Looking Glass Investments[30]

[16] https://waymo.com/mission/.

[17] https://www.nissanusa.com/experience-nissan/news-and-events/self-driving-autonomous-car.html.

[18] https://corporate.ford.com/articles/autonomous-technology/autonomous-2021.html.

[19] https://www.google.com/search/howsearchworks/algorithms/.

[20] https://help.twitter.com/en/using-twitter/twitter-timeline.

[21] https://www.spotify.com/.

[22] https://www.netflix.com.

[23] https://www.pressassociation.com/radarwebinar/..

[24] https://www.ibm.com/case-studies/ibm-watson-beat.

[25] https://melodrive.com/.

[26] https://www.jukedeck.com/.

[27] https://textio.com/.

[28] https://www.kavout.com/.

[29] https://greenkeytech.com/.

[30] https://www.lgiresearch.com/.

(also see Crosman 2017). Another widespread and long used application of AI in this area has been fraud detection, prevention, and management, provided by companies such as Feedzai[31] or FICO.[32] Lastly, AI is increasingly used to determine whether a prospective customer is eligible to receive a loan, an insurance, or a mortgage. Some exemplary businesses working on this type of technology are Experian,[33] PayPal,[34] and AliPay.[35] AI has proven itself incredibly useful in performing complex analysis and predictions based on large volumes of data, as is usually the case in the financial sector. However, using AI to perform high-speed transactions on the stock market can also be risky as shown by the 2010 flash crash experienced in the stock trading market (Keller 2012). Another potentially problematic development is exemplified by the Chinese Social Credit System, monitoring and rating every citizen's societal contribution and compliance to rules and governance, and determining their eligibility for schooling, travelling, matchmaking, loans, housing, jobs, licenses, visas, internet speeds, lower tax rates, public funding, investments, and more (Balistreri 2018; Botsman 2017; Kobie 2019; Ma 2018). Such an elaborate governmental monitoring system rewarding good and punishing bad behaviour through social status impact is at its best a highly effective mass behaviour manipulation system. Yet due to its complex and chaotic nature, feedback loops may occur, creating self-reinforcing downward spirals due to butterfly effects. This would cause potentially unfair decision-making towards individual citizens' social status and corresponding opportunities in life.

3.1.6 Logistics and decision support

Another example of AI development can be found in logistics. More and more companies use intelligent systems to optimise the transfer of goods between locations, and often, the actors performing the transfers are humans. Examples are Uber[36] and Deliveroo[37]—both of which gained bad publicity due to recent protests by employees who found themselves being exploited by AI-based scheduling algorithms (Reilly 2018). Yet other examples include Waze[38] and TomTom,[39] who spread traffic across the infrastructure so as to optimise time of arrival for all vehicles, yet—in doing so—also greatly disrupt large parts of the communal infrastructures (Madrigal 2018; Thai et al. 2016; Weise 2017). Impressive results obtained by AI applications can also be found in Decision Support. Nowadays, doctors are supported by algorithms able to recognise breast cancer (Bresnick 2017). For instance, InferVision[40] and Zebra Profound[41] both offer services to analyse CT and MRI images to recognise anomalies in patients' health and bodies (also see Ali 2019). And MedTelligent[42] and MatrixCare[43] provide healthcare management platforms connecting patients and doctors and offering all kinds of AI-based analytics that aim to improve personalised care through accurate diagnosis and treatment. DeepMind[44] as well as IBM Watson[45] are being used in healthcare, for instance to discover new medicines, or ensure that professionals have access to the right (secure) streams of (patient) information. Another area where decision support based in AI is on the rise is Human Resources. For example, HireVue[46] supports the prediction of performance for newly recruited talent, and MontageTalent[47] also promises to offer "a high-tech hiring experience for the modern candidate". Within the safety and security domain, the use of decision support systems to, for example, predict what city areas are most prone to car thefts or burglaries, and offer suggestions for additional patrolling (Smit et al. 2016). The most challenging risks within this field can be roughly placed into two categories that may reinforce one another. The first issue refers to emergent behaviour at the systemic level, as can be seen in the traffic example offered in the above. Emergence presents itself in other areas as well, for instance in cases where optimization at the individual level is not necessarily beneficial at the group or societal level, e.g. when only hiring people with construction skills for a construction company, or when consistently recommending the same treatment for specific cases within medicine. The second issue is bias, either in the dataset or in the model underlying the dataset. Examples are the hiring of men for top positions because historical data suggest that in the past men were successful in such positions (Dastin 2018), or the underrepresentation of women in healthcare studies resulting in treatment plans overfitted to the male population (Pressler 2016; Liu and Mager 2016). More generally speaking, applications of AI in the domain of logistics and decision support gradually shift the responsibility and oversight of large

[31] https://feedzai.com/.

[32] https://www.fico.com/.

[33] https://www.experian.com/.

[34] https://www.paypal.com/nl/home.

[35] https://intl.alipay.com/.

[36] https://uber.com.

[37] https://deliveroo.com.

[38] https://waze.com.

[39] https://tomtom.com.

[40] https://www.infervision.com/en.

[41] https://www.zebra-med.com/solutions/.

[42] https://www.medtelligent.com/.

[43] https://www.matrixcare.com/.

[44] https://deepmind.com/applied/deepmind-health/.

[45] https://www.ibm.com/watson/health/.

[46] https://www.hirevue.com/.

[47] https://www.montagetalent.com/.

socio-technical systems away from human planners and decision makers and place it in the hands of AI algorithms, causing emergent and biased effects at the systemic level that are hard to predict, understand, and control, and that are—at times—only uncovered after a major societal disruption.

3.1.7 Swarms and Robots

The previous examples were all mainly virtual applications, although, e.g. Automated Driving, Logistics AI, and Decision-Making AI are all strongly connected to very physical activities. One of the obvious physical AI applications is the domain of embodied intelligence, where robots, (swarming) drones, and intelligent weapon systems proliferate. Well-known impressive examples are the robots created by Boston Dynamics,[48] such as Atlas, which can perform a backflip, and SpotMini, which is capable of opening doors. Other accomplishments are obtained by numerous teams competing in, for instance, RoboCup Soccer.[49] Military applications have long integrated AI in their (semi-)autonomous platforms. Developed already in the 1960s and 1970s, currently used weapon systems, such as the MIM-104 Patriot Air Defense System[50] and the Goalkeeper,[51] are capable of autonomously searching, detecting, tracking, and taking out incoming missiles. State-of-the-art unmanned combat aerial vehicles, such as the Northrop–Grumman X47-B,[52] BAE Systems Taranis,[53] and Dassault nEUROn,[54] are autonomously capable of taking off, navigating, landing, mid-flight refuelling, evasive manoeuvring, and target identification (Ekelhof 2018). Potentially most notable are systems such as the IAI Harpy,[55] so-called 'kamikaze drones' or 'loitering munitions', that carry a high explosive warhead, and are capable of identifying and attacking a target, e.g. a radar emitter, all by itself (Simonite 2017; Winter 2017). Many military systems, however, also include possibilities for humans to stay "in the loop", allowing them to monitor and control the behaviour of the system or to intervene in cases where the system no longer behaves in line with human intent, military laws, or rules of engagement.

The military domain is not the only place where robot technology is on the rise. In the medical domain, one might run into social robots, such as Aldebaran/Softbank's Pepper,[56] TinyBots' Tessa,[57] or AIST's PARO robot.[58] Yet specialised surgical robots, such as the Da Vinci robot[59] or CMR Surgical's Versius,[60] can also be found. Within Urban Search and Rescue, the use of robots has made some major developments as well, using robots to detect and rescue survivors in the rubble (Davids 2002). And in warehouses and factories, companies such as Fetch Robotics,[61] Prime Robotics,[62] Bleum,[63] Total Productivity,[64] or Kuka Robotics[65] offer robotic solutions leading to improvements in speed, safety, accuracy, and customer satisfaction.

3.2 Analysis and reflection

Looking at the entire body of AI applications discussed in the above, evidence can be found supporting as well as refuting the tenets associated with each of the three perspectives on (the future of) AI, as reported in the following subsections.

3.2.1 Technology-centric perspective

The supporting and refuting evidence for the technology-centric perspective that can be distilled from the previous discussion is summarised in Table 1.

The main point that can be concluded from Table 1 is that the arguments presented for and against techno-centrism depend on the type of applications under consideration. As noted earlier, the technology-centric perspective fits the developments in game-playing AI. Many games have long been found to be huge challenges for any computer program to beat due to their computational complexity. However, their achievements all apply to finite games played within relatively closed game environments. The real world is more resembling of an infinite game (Carse 2011). An infinite game is characterized by its multi-player nature, the goal to achieve a diverse set of aims, dynamically defined rules that evolve through agreement of the participants, a lack of a clear-cut division between winners and losers. Many of the supporting evidence for the technology-centric perspective apply to finite games, while many of the refuting evidence

[48] https://www.bostondynamics.com/.

[49] https://www.robocup.org/.

[50] https://en.wikipedia.org/wiki/MIM-104_Patriot.

[51] https://en.wikipedia.org/wiki/Goalkeeper_CIWS.

[52] https://en.wikipedia.org/wiki/Northrop_Grumman_X-47B.

[53] https://en.wikipedia.org/wiki/BAE_Systems_Taranis.

[54] https://en.wikipedia.org/wiki/Dassault_nEUROn.

[55] https://en.wikipedia.org/wiki/IAI_Harpy.

[56] https://www.softbankrobotics.com/emea/en.

[57] https://www.tinybots.nl/.

[58] https://www.parorobots.com/.

[59] https://www.davincisurgery.com/.

[60] https://cmrsurgical.com/versius/.

[61] https://fetchrobotics.com/.

[62] https://www.primerobotics.com/.

[63] https://www.bleum.com/warehouse-robotics/.

[64] https://totalproductivity.nl/en/products/industrial-robots/.

[65] https://www.kuka.com/.

Table 1 Supporting and refuting evidence for the technology-centric perspective

	Supporting evidence	Refuting evidence
T1 When sufficiently developed, AI technology can be applied to solve any problem	Recent AI progress has led to game AI, highly autonomous cars, FinTech applications, etc. that were inconceivable a decade ago	New AI applications in healthcare or finance decision support systems, do not solve a problem by themselves, but support the human in problem solving
T2 AI technology may introduce additional problems which can, in turn, be solved by AI	The Fintech flash crash in 2010, and the Twitterbot Tay problem in 2016 were one-off incidents, and should be regarded as teething problems	Problems with filter bubbles, behaviour modification algorithms, deep fakes, have only increased despite significant efforts of finding a technological solution
T3 The more AI is developed, the less user interaction is needed	Alpha Go Zero learned to beat the human Go champion just by playing against itself. No user interaction needed	A strong need for explainable AI has emerged in AI decision support systems indicating that user interaction will change rather than become obsolete
T4 Current AI technology has only reached a fraction of its full development potential	AlphaZero, DeepMind, and Watson are reusable algorithms capable of taking on a wider variety of challenges	Many of the widely applied algorithms have been around for half a century. Any improvements in their application nowadays comes from better understanding, improvements in usability and distribution, or higher availability of training data
T5 Artificial Intelligence has vastly more potential than human intelligence	Over the last decade AI has steadily been achieving superhuman performance in, e.g. game AI, pattern recognition. This trend can be expected to continue	When looking at AI systems in all application domains, we observe that even though AI systems are capable of outsmarting humans in certain parts of the task execution (e.g. parking, staying within one's driving lane, recognising malignant cancerous tissues, detecting anomalies in financial transactions, and so on), none of these AI systems is capable of performing the whole task. Humans are still needed to, e.g. offer additional interpretation, handle rare events, or combine information coming from different (often also analogue) sources

Table 2 Supporting and refuting evidence for the human-centric perspective

	Supporting evidence	Refuting evidence
H1 Artificial intelligence only exhibits part of human cognition and is therefore insufficient for many real-world problems	When looking at the examples of conversational agents, decision support, and autonomous car technology, one can observe the limits of AI's performance in complex dynamic environments	Recent achievements of artificial intelligence in tasks traditionally considered hard (e.g. boardgames, medicine, fintech, logistics, etc.) show that AI can outsmart humans
H2 Artificial intelligence will remain relatively limited for the foreseeable future	AI degrades rapidly in the face of unexpected and/or unknown situations. As can be seen from the examples, autonomous driving within busy environments, or upholding a complex conversation are—as we speak—tasks difficult to difficult to accomplish with AI	Artificial intelligence is present in almost every area of society, facilitating important decisions, speeding up processes, enabling more efficient and effective performance in areas such as logistics, health care, and insurance
H3 Problems that are caused by AI cannot be solved by applying more AI	AI lacks the contextual understanding required to distinguish true (cor)relations from coincidental ones (Amazon's recruitment algorithm), or socially acceptable input from intentionally disrespectful input (Microsoft Tay). Humans can rely on their common-sense capabilities to "fix" models by careful interpretation, analysis, hypothesising, testing, and restructurin	New features added to, e.g. decision support systems, allow for humans to inspect the reasoning behind the suggestions made by the software, and notify human decision makers of potential biases
H4 AI technology often introduces additional problems for human well-being, which should be a reason to rethink whether the technology should be applied	Examples such as Deliveroo and Uber exploiting its employees, or Waze, TomTom and Google Maps derailing local infrastructure, or Amazon's sexist recruiting algorithm show how detrimental the effects of AI can be	Examples, such as the analysis of healthcare imagery to detect cancerous cells and other health risks, show that the world is better off with AI aiding people to provide better care, and so we should continue to do so
H5 Artificial intelligence is useful for supporting humans and will never act without human involvement	Self-driving cars still remain at level 3, supporting human drivers, and requiring them to intervene or take over under certain conditions. Personal conversational assistants are aimed at supporting humans, and rely on many human-produced webservices, such as Wikipedia or Weather. Creative content and media AI, as well as logistics and decision support systems, rely on large volumes of human-produced data	Game-playing AI has taught itself to play, following minimal rules and domain knowledge provided by humans

Table 3 Supporting and refuting evidence for the collective-intelligence perspective

	Supporting evidence	Refuting evidence
C1 Intelligence should not be studied at the level of individual humans or AI machines, but at the group level of humans and AI machines working together	In various domains, unforeseen emergent effects at the systemic level can be observed, such as with Deliveroo, Uber, Waze, and Google maps, or with hiring software and other decision support systems	Part of the problems seen in AI systems can in fact be understood and solved at the level of the individual AI system, such as bugs, flawed algorithms, missing domain knowledge, or erroneous reasoning rules
C2 Increasing the intelligence of a system should be achieved by increasing the quality of the interaction between its constituents rather than the intelligence of the constituents themselves	Semi-autonomous cars and robots (e.g. UxVs) are as of yet uncapable of performing autonomously, but can outperform humans within certain situations. Currently, one of the biggest challenges in these fields is how to seamlessly integrate such systems in human processes and workflows	Past developments in self-driving car technology and robot technology are largely due to advancement in the car, cq. robot, technology and not so much improvement of human-system interaction or collaboration
C3 Both human as well as artificial intelligence can be regarded as very shallow when considered in isolation	When looking at Wikipedia, and other large collaborative platforms, like Uber and Deliveroo, it is easy to see that the whole is larger than the sum of its parts	Exceptions to the rule show that this is not always the case. People like Albert Einstein or Stephen Hawking have accomplished great work individually. And AlphaZero is capable of teaching and learning all by itself
C4 No AI is an island	Personal conversational assistants, for example, rely on many other webservices to create value for their customers. They are really networked systems of many different distributors and manufacturers	Idem

applies to infinite games required to effectively act in the real world.

3.2.2 Human-centric perspective

The supporting and refuting evidence for the human-centric perspective that can be distilled from the previous discussion is summarised in Table 2.

The main take-away from the findings listed in Table 2 is that this perspective is founded mostly in applications dealing with large, ill-structured, complex, and dynamic environments and a large set of integrated tasks and behaviours. In other words, the focus of the arguments presented by this perspective lies within a different segment of AI applications as compared to the focus used by the techno-centric perspective. The type of applications that drive the arguments of the human-centric perspective often require a tremendous effort in modelling relevant parts of the world, refining software and its problem-solving capabilities, designing reward functions of the AI, or selecting the right data to feed the models required for the AI to function properly. As a result, humans are indispensable in the design, development, and deployment of these systems. As a result, these systems are susceptible to the subjective values, needs, and interests of the people providing the necessary input. Examples of negative societal effects can be found in applications such as Deliveroo, AirBnB, and Uber (exploited employees), Waze and Google Maps (disrupted local infrastructure), Facebook and Instagram (clandestine advertisement companies), and the voice assistants created by Amazon, Apple, and Google (impulsive shopping behaviour, increased debt, and societal inequality).

3.2.3 Collective-Intelligence perspective

The supporting and refuting evidence for the human-centric perspective that can be distilled from the previous discussion is summarised in Table 3.

Table 3 shows that the scale at which AI is now distributed, multiplied, adapted, and vastly interconnected allows this technology to generate massive impact on society, at a rate that no longer allows for careful consideration of future consequences. As a result, the effects of changes to existing AI applications, or additions of new AI applications can quickly propagate throughout the networks with which they interconnect, thereby affecting large human organisations, infrastructures, companies, families, and other social structures across the globe. Such emergent effects are readily observable when looking at the effects of social media on politics, traffic obstructions caused by traffic routing applications, and the proliferation of giant enterprises—e.g. "the Big Five": Alphabet, Facebook, Apple, Microsoft, and Amazon—at the cost of smaller ones. These effects are amplified by the recent creation of a virtual world (i.e. the internet) in addition to the real one, a world that played no part in the evolution of the human body, and its sensory-motor capabilities nor its intelligence. Human intelligence is well equipped to deal with the physical world, the reality in which it was formed and trained. However, the virtual world now created is for a large part opaque and difficult to understand and predict for the human brain, and so people often succumb to anthropomorphism. Especially now that the virtual world, artificial intelligence, and networked information are so intertwined with the physical world, the challenges to harmonise human beings and their intelligence with the virtual world created by them must be addressed sooner rather than later.

3.3 Reflection upon the three perspectives

After this analysis of the three perspectives, the question that rises is: how do these three perspectives relate to each other? When looking at the different perceptions and the corresponding tenets, and the evidence that can be found for and against each of them, we observe that different arguments come from different domains and applications. A possible interpretation of the different perspectives may well be that each of the perspectives tends to focus on different achievements and within different application areas, resulting in the perception of different types of challenges that require different types of solutions. If this is indeed the case, then choosing for one or the other would lead to the overlooking of a large part of the application space, along with the corresponding achievements, challenges, and risks. When combining the three perspectives, a broader view on AI developments emerges, along with the possibility to observe effects that propagate through the entire application space as well.

Additionally, following the observations in the above, artificial intelligence and human intelligence should not be compared along the same dimension, a view also expressed by Dickson (2018). What is perceived as intelligent behaviour depends on the type of task and context. For the moment, human capabilities fundamentally differ from AI capabilities, as discussed in the previous subsections. Even so, debaters on the topic of the effects of AI for human society frequently fall into the trap of comparing human intelligence to artificial intelligence.

Moving on, many of the apparent disagreements between the perspectives stem from a different level of abstraction (in size or time) at which the system is regarded. For example, a robotic AI system might seem to explore an environment fully autonomously without human involvement (an argument seemingly coherent with the AI-techno-centric view). But there has always been human involvement prior to this phase, when the system was tasked to do so. Additionally, there is frequently a larger organisational structure that requires this task to be done for a greater purpose, almost always involving humans.

Lastly, humans and AI make decisions in different ways and it is, therefore, not appropriate to juxtapose them as mutually excluding. Instead, humans and AI should be seen as team members with different, largely complementary capabilities. A proper approach to AI engineering should regard intelligence on multiple abstraction levels, ranging from the individual, to the team, and society level. A serious challenge of the CI perspective is how to make them function in a synergistic way, and how to disentangle the various components and their effects so as to accommodate changes to the design and mitigate unwanted effects at the systemic level. To address this challenge, it is not enough to just consider human–AI teams, but it is necessary to look at the broader context in which systems function, including the implications they have for society as a whole. This broad perspective may be too complex for some purposes, in which an approach from techno- or human perspective could be more appropriate.

Summarising, we do not have to choose between any of the three views expressed above. However, we will aim for an artificial intelligence design method that allows us not to get trapped into the limitations of the above views. Therefore, we will use the term hybrid collective intelligence design method to denote the view that combines the best practices from each perspective. This approach for designing hybrid collective intelligence is a harmonious merger of the three perspectives described above. Depending on purpose and context, each of the perspectives of AI (techno-centric, human-centric, and collective intelligence) has its merits. What perspective is chosen to tackle a problem often depends on the personal conviction of the company, instead of a solid analysis. Scientists should be better equipped to decide which perspective is appropriate to study any given situation.

4 Developing hybrid collective intelligence

From the analysis presented in Sect. 3, we conclude that each of the three perspectives has important added value when developing AI systems. However, the perspective that is used as an underlying system development philosophy is often not a deliberate choice, but a coincidental matter of who happens to be in the development team. Depending on their background, scholars are often naturally drawn towards one of the three perspectives. The merits of each perspective can only be achieved within a diverse development team in which different opinions are respected and equally valued.

Whereas all perspectives would ascribe to the idea that humans always remain involved, they would disagree on the phase at which humans would be involved. For the

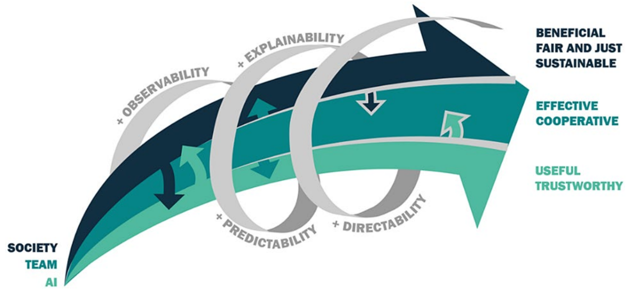


Fig. 4 Development of collective intelligence

techno-centrists, humans are primarily involved in the design and development phases during which engineers and programmers build and train AI technology. For the human-centrists, however, humans not only build the AI, but also they remain important afterwards to warn against AI overlooking humankind's social nature in technological designs, and to interact with the AI throughout its operational lifecycle. Within the collective intelligence-centric perspective, engineers may collaborate with users during construction, implementation, and in everyday practices, as—according to this view—true intelligence is regarded to be seated in the collective of intelligent human and artificial agents.

In the following, we propose a set of methodological design principles that lead to the combination of the three perspectives. The aim of these principles is to promote that a problem is tackled from the right perspective(s) during each phase of development and deployment. In general, the appropriateness of each perspective depends on the particular design objective one wishes to pursue.

By viewing the AI system as the locus of intelligence (at least once it is designed, programmed, and trained by human engineers), the technocentric perspective is well suited to design an AI system that performs well in terms of *system performance measures*, such as *classification accuracy*, *stability,* and *speed*. The humans who designed, programmed and trained the system, however, should still aim to foster human wellbeing and have knowledge about how to accomplish this.

By viewing the human as the locus of intelligence, the human-centric perspective is well suited to design an AI system that *interacts with humans to foster human wellbeing*. To accomplish this, designers and developers should consider the well-being of people other than themselves while designing, programming, and training their AI system. It is essential to consider the humans who are ultimately exposed to the AI system's behaviour and its effects. This approach can be recognised in modern system design approaches, such as value sensitive design (Friedman et al. 2013), privacy by design (Cavoukian 2013), secure by design (Santos et al. 2017), and others. By limiting an AI system's autonomous capabilities to the bare minimum needed for it to achieve the desired level of performance, predictable and controllable behaviour can be warranted as much as possible. This may help prevent AI systems from depriving humans of their *sense of control* and helps foster humans' capacity to be *resilient* and compensate for the machine's weaknesses.

By viewing the collective of humans and machines as the locus of intelligence, the collective intelligence perspective is well suited to analyse and design AI systems that perform well on *properties that emerge on a macro scale*. Examples of these are *equality*, *fairness*, and *sustainability,* all values emerging on a societal level. None of these properties can be attributed to one single human or AI component. The CI perspective is also useful to pursue goals that are emergent on a smaller group level, such as *team resilience* (i.e. the capacity of team members to take over each other's work when one component breaks down). The gathering of human–AI teams at the level of society may also introduce challenges at the level of the societal eco-system, that are difficult to grasp from a techno-centric or human-centric perspective. Such challenges include misunderstanding or misalignment between stakeholder groups resulting in unintentional injustice or discrimination, but it could also concern deliberate obfuscation or wrongdoing by one group to achieve an advantage over another group. An example is commercial motivation leading a firm to seek for the obtainment of economic dominance—this may result in advantage for the subgroup, but may lead to unfavourable effects on the larger group. Both positive and negative effects play at the collective level, and demand analysis from a CI perspective.

As argued in the previous section, the trend that AI is becoming more networked and connected leads to a higher importance of adopting the collective intelligence perspective. Nevertheless, the design of AI systems should push towards the achievement of objectives established at the collective level *in addition* to the accomplishment of local goals, such as technological achievement or human-centric performance. How to translate the three perspectives into a coherent design methodology can be regarded as one of the major challenges for the coming decades. A multi-level view on the effects of AI is a large research field on its own, and the scientific community has barely begun to scratch its surface (Rahwan et al. 2019). From a design perspective, it can be noted that developments should proceed in three strands (depicted in Fig. 4):

1. At the level of the **AI** application, the development will be done mostly from a techno-centric perspective. By this, we do not mean to say that all developers will endorse all tenets of techno-centrism as presented in Fig. 1, but the developers will regard improvements to the AI system as the main way to enhance intelligent behaviour. This is simply because the system boundary

does not extend beyond the technical AI system. For this purpose, the classic cycle of requirements engineering, prototyping, and evaluation is iteratively performed to develop a system that is compliant with important goals at the system level. These could, for example, be reliability, speed, and sufficient performance on a test dataset. Once the system is up and running, it may develop itself further by learning from new training data, so as to improve its behaviour and performance regarding said values. Selection, label, and inductive bias must be addressed during this *continual learning* process. Involving humans in collaboration loops with AI requires, for example, for the joint human–AI system to be aware of bias (e.g. through the addition of smart feedback loops), and requires for the implementation of methods for detecting and (if possible) remedying bias.

2. At the level of **teams**, AI applications and humans together form human–agent teams (HAT) (Johnson and Vera 2019) capable of performing tasks in an effective and cooperative way. HATs are developed, for example, by designing appropriate interactive behaviours for the AI applications, and by providing appropriate training to the human team members. During the operational phase, the HAT will develop itself further towards to-be-defined values important at the team level, such as effective and cooperative team behaviour (van Diggelen et al. 2018; de Visser et al. 2019). At this level, values that are typically put forward by human-centrism and the collective perspective can be addressed. Appropriate human team bias assessment must take place, in addition to the inductive bias, label and selection assessment in the first strand (also see HumBL, 2019).
3. At the level of **society**, the different HATs are assembled to form a more or less coherent community or ecosystem. At the society level, there are also important values that can guide society as a whole towards optimal performance. Examples may be to optimise towards a beneficial, fair, and just, or perhaps sustainable systemic interconnectedness. Typically, these are studied from a collective or human-centric perspective.

The principles for designing hybrid collective intelligence can be summarised as follows:

Design principle 1 AI system design must simultaneously consider goals from a collective intelligence, techno-centric and human-centric perspective.

A major challenge is to design an AI system with the reciprocal relation between society and its human and machine members in mind. In Fig. 4, these interdependencies are depicted using the vertical arrows between the three lines of development. To predict the effects of AI, one needs to intimately understand how AI systems, humans, and society relate. As they are a member of society, AI entities can change the culture of the society, which in turn changes the data they feed on, and hence their own behaviours. This reciprocal relation between society and its human and machine members is extremely complicated, and most likely will always involve a certain degree of uncertainty. We can aim for a design method that minimises undesirable consequences of AI, but these can never be fully avoided. This is particularly true for AI systems that are placed in a context upon which they are heavily dependent, but which is not known at design time. It is also true for learning systems that change their behaviour based on training data they encounter at runtime (such as Twitterbot Tay). Therefore, we argue that ensuring desired AI behaviour does not end after the design phase but remains a continuous effort over the entire lifecycle of a product.

Design principle 2 Pursuing design objectives of AI systems demands a continuous effort over the entire lifecycle of a product.

To allow actors to spend this effort, they must be aware of the current situation, where it is heading and how they can change it. This requires a continuous process of observing, predicting, explaining, and directing by all constituents in the Human–AI Society. This principle is depicted in Fig. 4 as the spiral around all three levels of design. We identify four important requirements for the effective design of collective intelligence: Observability, Predictability, Explainability, and Directability (OPED). The requirements for OPD have been proposed by (Johnson et al. 2014) as the main high-level requirements for human agent teamwork. Observability means that an actor should make its status, its knowledge of the team, task, and environment observable to others. Predictability means that an actor should behave predictably such that others can rely on them when considering their own actions. Directability means that actors should have the opportunity to (re-)direct each other's behaviour. We add Explainability to this list, which means that agents should be capable of explaining their behaviour to others (Neerincx et al. 2018).

Design principle 3 AI must be developed in a way that provides observability, predictability, explainability, and directability at all abstraction levels (AI, team, and society).

The requirements for OPED apply to all three levels of design (AI, Team, Society). This leads to twelve combinations that must, in some way, be satisfied. For example, consider a loan assessment AI system as described in Sect. 3.1.5. Explainability at the AI level may involve the system explaining to its loan-applicant why it has denied a certain application. Observability by the same system could involve a way of making the user aware that the system is currently processing a request. Observability at the society level can be recognised in a journalism organisation such as Pro Publica,[66], [67] that monitors AI-based applications for social injustice, such as discrimination against certain minority groups. The directability at the societal level could be established by drawing the public's attention to the matter using journalism. An example of a research project directed at fostering observability, predictability, explainability, and directability can be found in an EU Horizon 2020 project called REELER,[68] where a new type of intermediaries, called "alignment experts" are responsible for aligning the values of different stakeholders and use the outcomes as input for the design of an AI system (also see https://responsiblerobotics.org/). Whereas these examples show that different mechanisms are already arising in society, they do so in an uncoordinated way. We argue that they should be an integral part of the design of AI systems.

5 Conclusion

Debates about (future) effects of AI on human society are dominated by three perspectives: the techno-centric perspective, the human-centric perspective, and the collective intelligence centric perspective. In this paper, we showed that each of the three perspectives offers a unique contribution to the debate resulting from their differences in focus and background knowledge in specific applications and corresponding opportunities, risks, and challenges. Combining the three perspectives into a single integrated and comprehensive framework allows for researchers and developers to adopt an appropriate perspective when tackling a given design challenge. This framework fosters a 360º view on the entire problem and solution space. Such a wide-angle view allows researchers and designers to reach a better understanding of how design choices made when thinking and working from one perspective affect phenomena studied and observed, or effects identified as risky or fruitful, by another perspective. We provided three design principles to accommodate this holistic view on the future of AI research, design, and development. Future research will aim to further expand the framework, its design principles, and will deliver additional design methods to accommodate a wide perspective on AI research, design, and development, harnessing the strength of each of the three perspectives.

Acknowledgments This material is based upon work supported by the Dutch Ministry of Defence's exploratory research program (project Human-AI Teaming).

References

Adamson G, Havens JC, Chatila R (2019) Designing a value-driven future for ethical autonomous and intelligent systems. Proc IEEE 107(3):518–525

Ali AR (2019) Deep learning in oncology—applications in fighting cancer. EMERJ. https://emerj.com/ai-sector-overviews/deep-learning-in-oncology/. Accessed 26 Feb 2019

Ashby WR (1961) An introduction to cybernetics. Chapman & Hall Ltd

Awad E, Dsouza S, Bonnefon JF, Shariff A, Rahwan I (2020) Crowdsourcing moral machines. Commun ACM 63(3):48–55

Bailey I, Wilson GA (2009) Theorising transitional pathways in response to climate change: technocentrism, ecocentrism, and the carbon economy. Environ Plan 41(10):2324–2341

Balistreri H (2018) Credit as a social technology: black mirror, China, and the case for social credit. Medium. https://medium.com/@harrison.tb/credit-as-a-social-technology-black-mirror-china-and-the-case-for-social-credit-d6a6db609ea5. Accessed 17 Jun 2020

Basem W (2018) Audi A8 L review|brilliant engineering in an unassuming wrapper. https://www.autoblog.com/2018/10/16/2019-audi-a8-l-review-first-drive/. Accessed 3 Apr 2019

Baum SD (2017) On the promotion of safe and socially beneficial artificial intelligence. AI & Soc 32(4):543–551

Bostrom N (2016) Superintelligence. Oxford University Press, Oxford

Bostrom N, Yudkowsky E (2014) The ethics of artificial intelligence. Camb Handb Artif Intell 316:334

Botsman R (2017) Big data meets Big Brother as China moves to rate its citizens. https://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion. Accessed 17 Jun 2020

Bresnick J (2017) Deep learning network 100% accurate at identifying breast cancer. https://healthitanalytics.com/news/deep-learning-network-100-accurate-at-identifying-breast-cancer. Accessed 28 Mar 2019

Brown N, Sandholm T (2018) Superhuman AI for heads-up no-limit poker: libratus beats top professionals. Science 359(6374):418–424. https://doi.org/10.1126/science.aao1733

[66] https://www.propublica.org/.

[67] https://www.yonder-ai.com/.

[68] https://reeler.eu/.

Brynjolfsson E, McAfee A (2014) The second machine age: work, progress, and prosperity in a time of brilliant technologies. WW Norton & Company, New York

Brynjolfsson E, Mitchell T (2017) What can machine learning do? Workforce implications. Science 358(6370):1530–1534

Bughin J, Hazan E, Ramaswamy S, Chui M, Allas T, Dahlström P, Henke N, Trench M (2017) Artificial Intelligence: the next digital frontier?. McKinsey Global Institute, New York

Campbell M, Hoane AJ Jr, Hsu FH (2002) Deep blue. Artif Intell 134(1–2):57–83

Carse J (2011) Finite and infinite games. Simon and Schuster, New York

Case N (2018) How to become a centaur. Journal of Design and Science

Cavoukian A (2013) Privacy by design: leadership, methods, and results. In: European Data Protection: Coming of Age (pp. 175-202). Springer, Dordrecht

Chen Y, Elenee Argentinis J, Weber G (2016) IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research. Clin Ther 38(4):688–701. https://doi.org/10.1016/j.clinthera.2015.12.001

Clark C, Chalmers D (1998) The extended mind. Analysis 58(1):7–19

Constine J (2016) How Facebook news feed works. http://social.techcrunch.com/2016/09/06/ultimate-guide-to-the-news-feed/. Accessed 28 Mar 2019

Crosman P (2017) Beyond robo-advisers: how AI could rewire wealth management. https://www.americanbanker.com/news/beyond-robo-advisers-how-ai-could-rewire-wealth-management. Accessed 28 Feb 2019

Dar P (2018) Major AI and ML breakthroughs in 2018 and trends to look out for in 2019. https://www.analyticsvidhya.com/blog/2018/12/key-breakthroughs-ai-ml-2018-trends-2019/. Accessed 27 Mar 2019

Dastin J (2018) Amazon scraps secret AI recruiting tool that showed bias against women. Reuters Business News. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. Accessed 6 Jan 2020

Davids A (2002) Urban search and rescue robots: from tragedy to technology. IEEE Intell Syst 17(2):81–83

Davies A (2018) What is a self-driving car? The complete WIRED guide. Wired. https://www.wired.com/story/guide-self-driving-cars/. Accessed 17 Jun 2020

de Visser EJ, Peeters MMM, Jung MF, Kohn S, Shaw TH, Pak R, Neerincx MA (2019) Towards a theory of longitudinal trust calibration in human-robot teams. Int J Soc Robot. https://doi.org/10.1007/s12369-019-00596-x

de Winter JC, Dodou D (2014) Why the Fitts list has persisted throughout the history of function allocation. Cogn Technol Work 16(1):1–11

Deahl D (2018) Google's NSynth super is an AI-backed touchscreen synth. Verge. https://www.theverge.com/circuitbreaker/2018/3/13/17114760/google-nsynth-super-ai-touchscreen-synth. Accessed 17 Jun 2020

Dellermann D, Ebel P, Söllner M, Leimeister JM (2019) Hybrid intelligence. Bus Inf Syst Eng 61(5):637–643

Department of Defense (2015) Technical assessment: autonomy. Office of Technical Intelligence, Department of Defense, Washington, DC

Dickson B (2018) There's a huge difference between AI and human intelligence—so let's stop comparing them. Tech Talks Blog. https://bdtechtalks.com/2018/08/21/artificial-intelligence-vs-human-mind-brain/. Accessed 17 Jun 2020

Dorado GL, Blockwood JC, Persons TM, Sanford S (2018) Artificial intelligence—emerging opportunities, challenges, and implications. United States Government Accountability Office

Ekelhof, M. A. (2018). LIFTING THE FOG OF TARGETING. Naval War College Review 71(3):61-95

Endsley MR (2018) Level of automation forms a key aspect of autonomy design. J Cognit Eng Decis Mak 12(1):29–34

Engel D, Woolley AW, Jing LX, Chabris CF, Malone TW (2014) Reading the mind in the eyes or reading between the lines? Theory of mind predicts collective intelligence equally well online and face-to-face. PLoS ONE 9(12):e115212. https://doi.org/10.1371/journal.pone.0115212

Engelbart DC (1962) Augmenting human intellect: a conceptual framework. Stanford Research Institute, Menlo Park

Feltovich PJ, Prietula MJ, Anders Ericsson K (2006) Studies of expertise from psychological perspectives. In: Anders Ericsson K, Charness N, Feltovich PJ, Hoffman RR (eds) The Cambridge handbook of expertise and expert performance. Cambridge University Press, New York, pp 41–68

Ferrucci DA (2012) Introduction to "This is Watson". IBM J Res Dev 56(3.4):1:1–1:15. https://doi.org/10.1147/JRD.2012.2184356

Fitts PM (ed) (1951) Human engineering for an effective air navigation and traffic control system. National Research Council, Washington, DC

Ford M (2018) Architects of intelligence: the truth about AI from the people building it. Packt Publishing, Birmingham

Friedman B, Kahn PH, Borning A, Huldtgren A (2013) Value sensitive design and information systems. In: Early engagement and new technologies: opening up the laboratory. Springer, Dordrecht, pp 55–95

Future of Life Institute (FLI) (2015) Research priorities for robust and beneficial artificial intelligence: an open letter. http://futureoflife.org/AI/open_letter. Accessed 3 Aug 2015

González RJ (2017) Hacking the citizenry?: personality profiling, 'Big Data' and the election of donald trump. Anthropol Today 33:9–12. https://doi.org/10.1111/1467-8322.12348

Hadfield-Menell D, Dragan A, Abbeel P, Russell S (2017) The off-switch game. In: Workshop at the thirty-first AAAI conference on artificial intelligence. https://www.aaai.org/ocs/index.php/WS/AAAIW17/paper/view/15156. Accessed 17 Jun 2020

Hao K (2019) DeepMind wants to teach AI to play a card game that is harder than Go. MIT Technol Rev. https://www.technologyreview.com/s/612886/deepmind-wants-to-teach-ai-how-to-play-a-card-game-thats-harder-than-go/. Accessed 17 Jun 2020

Hawkins AJ (2018) Riding in Waymo One, the Google spinoff's first self-driving taxi service. Verge. https://www.theverge.com/2018/12/5/18126103/waymo-one-self-driving-taxi-service-ride-safety-alphabet-cost-app. Accessed 17 Jun 2020

Henrich J (2015) The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter. Princeton University Press, Princeton

High R (2012) The era of cognitive systems: an inside look at IBM Watson and how it works. IBM, New York, p 16

Hollnagel E, Woods DD (2005) Joint cognitive systems: foundations of cognitive systems engineering. CRC Press, Boca Raton

Holloway D, Green L (2016) The internet of toys. Commun Res Pract 2(4):506–519. https://doi.org/10.1080/22041451.2016.1266124

Horton H (2016) Microsoft deletes "teen girl" AI after it became a Hitler loving sex robot within 24 hours. Telegraph. https://www.telegraph.co.uk/technology/2016/03/24/microsofts-teen-girl-ai-turns-into-a-hitler-loving-sex-robot-wit/. Accessed 17 Jun 2020

Hüllermeier E, Fober T, Mernberger M (2013) Inductive bias. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H (eds) Encyclopaedia of systems biology. Springer, New York

HumBL: Bias-aware humans-in-the-loop (HumBL) workshops. https://humlworkshop.github.io/HumBL-WWW2019

Isaak J, Hanna MJ (2018) User data privacy: Facebook, Cambridge analytica, and privacy protection. Computer 51(8):56–59. https://doi.org/10.1109/MC.2018.3191268

Ito J (2019) Forget about artificial intelligence, extended intelligence is the future. Wired UK. https://www.wired.co.uk/article/artificial-intelligence-extended-intelligence. Accessed 17 Jun 2020

Jiang H, Nachum O (2019) Identifying and correcting label bias in machine learning. https://arxiv.org/abs/1901.04966

Johnson M, Vera A (2019) No AI is an island: the case for teaming intelligence. AI Mag 40(1):16–28

Johnson M, Bradshaw JM, Feltovich PJ, Jonker CM, van Riemsdijk MB, Sierhuis M (2014) Coactive design: designing support for interdependence in joint activity. J Hum-Rob t Interact 3(1):43–69

Kahneman D (2011) Thinking, fast and slow, vol 1. Farrar, Straus, and Giroux, New York

Kamar E (2016) Directions in hybrid intelligence: complementing AI systems with human intelligence. In: IJCAI, pp 4070–4073

Kamphorst BA (2012) The primacy of human autonomy: understanding agent rights through the human rights framework. In: Proceedings of the 1st workshop on rights and duties of autonomous agents (RDA2), vol 885, pp 19–24

Kamphorst B, Kalis A (2015) Why option generation matters for the design of autonomous e-coaching systems. AI & Soc 30(1):77–88

Keller AJ (2012) Robocops: regulating high frequency trading after the flash crash of 2010. Ohio State Law J 73:1457

Klein G, Shneiderman B, Hoffman RR, Wears RL (2020) The "war" on expertise: five communities that seek to discredit expertise. In: Ward P, Schraagen JM, Gore J, Roth E (eds) The Oxford handbook of expertise. Oxford University Press, Oxford, pp 1157–1192

Kobie N (2019) The complicated truth about China's social credit system. Wired UK. https://www.wired.co.uk/article/china-social-credit-system-explained. Accessed 17 Jun 2020

Kunze L, Hawes N, Duckett T, Hanheide M, Krajník T (2018) Artificial intelligence for long-term robot autonomy: a survey. ArXiv:http://arxiv.org/abs/1807.05196 [Cs]. http://arxiv.org/abs/1807.05196

Kurzweil R (2005) The singularity is near: when humans transcend biology. Penguin, New York

Lanier J (2018) Excerpt from "Ten arguments for deleting your social media accounts right now". Henry Holt. https://www.barnesandnoble.com/readouts/ten-arguments-for-deleting-your-social-media-accounts-right-now/. Accessed 26 May 2019

LeBeau P (2018) GM is seeking approval for an autonomous car that has no steering wheel or pedals. CNBC. https://www.cnbc.com/2018/01/12/gm-is-seeking-approval-for-an-autonomous-car-that-has-no-steering-wheel-or-pedals.html. Accessed 17 Jun 2020

Legg S, Hutter M (2007) A collection of definitions of intelligence. Front Artif Intell Appl 157:17

Licklider JC (1960) Man-computer symbiosis. IRE Trans Hum Fact Electron 1:4–11

Lin P (2013) The ethics of autonomous cars. Atlantic. https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/. Accessed 17 Jun 2020

Lin P (2014) The robot car of tomorrow may just be programmed to hit you. Wired. https://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you/. Accessed 17 Jun 2020

Lin P (2014) Here's a terrible idea: robot cars with adjustable ethics settings. Wired. https://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings/. Accessed 17 Jun 2020

Liu KA, Dipietro Mager NA (2016) Women's involvement in clinical trials: historical perspective and future implications. Pharmacy Practice (Granada) 14(1):

Lloyd K (2018) Bias amplification in artificial intelligence systems. https://arxiv.org/abs/1809.07842

López G, Quesada L, Guerrero LA (2018) Alexa vs. Siri vs. Cortana vs. Google assistant: a comparison of speech-based natural user interfaces. In: Nunes IL (ed) Advances in human factors and systems interaction. Springer International Publishing, Berlin, pp 241–250

Loucks J, Hupfer S, Jarvis D, Murphy T (2019) Future in the balance? How countries are pursuing an AI advantage [Deloitte Insights]. Deloitte, London

Ma A (2018) China has started ranking citizens with a creepy "social credit" system—here's what you can do wrong, and the embarrassing, demeaning ways they can punish you. https://www.businessinsider.com/china-social-credit-system-punishments-and-rewards-explained-2018-4. Accessed 17 Apr 2019

Madrigal AC (2018) The perfect selfishness of mapping apps. Atlantic. https://www.theatlantic.com/technology/archive/2018/03/mapping-apps-and-the-price-of-anarchy/555551/. Accessed 17 Jun 2020

Maheshwari S (2017) Burger king "Ok, Google" ad doesn't seem ok with Google. New York Times. https://www.nytimes.com/2017/04/12/business/burger-king-tv-ad-google-home.html. Accessed 17 Jun 2020

Malone TW (2018) How human-computer 'Superminds' are redefining the future of work. MIT Sloan Manag Rev 59(4):33–42

Malone TW, Bernstein MS (2015) Handbook of collective intelligence. MIT Press, London

Maurer M, Gerdes JC, Lenz B, Winner H (2016) Autonomous driving—technical, legal and social aspects. Springer, Berlin

Metz C (2016) How Google's AI viewed the move no human could understand. https://www.wired.com/2016/03/googles-ai-viewed-move-no-human-understand/. Accessed 20 Mar 2019

Metz C (2018) How will we outsmart A.I. liars? New York Times. https://www.nytimes.com/2018/11/19/science/artificial-intelligence-deepfakes-fake-news.html. Accessed 17 June 2020

Miller AP (2018) Want less-biased decisions? Use algorithms. Harvard business review, 2.

Mittrick M, Richardson J, Dennison Jr M, Trout T, Heilman E, Hanratty T (2019) Investigating immersive collective intelligence. In: Artificial intelligence and machine learning for multi-domain operations applications. International Society for Optics and Photonics, vol 11006

Morgan S (2018) Fake news, disinformation, manipulation, and online tactics to undermine democracy. J Cyber Policy 3(1):39–43

Mulgan G (2017) Big mind: how collective intelligence can change our world. Princeton University Press, Princeton

Neerincx MA, van der Waa J, Kaptein F, van Diggelen J (2018) Using perceptual and cognitive explanations for enhanced human-agent team performance. In: International conference on engineering psychology and cognitive ergonomics. Springer, Cham, pp 204–214

Newton-Rex E (2017) 59 impressive things artificial intelligence can do today. https://www.businessinsider.com/artificial-intelligence-ai-most-impressive-achievements-2017-3. Accessed 27 Mar 2019

Ng, A. (2016). What artificial intelligence can and can't do right now. Harvard Business Review, 9.

O'Neil C (2017) Weapons of math destruction: how big data increases inequality and threatens democracy. Broadway Books, New York

Osoba OA., Welser W (2017). The risks of artificial intelligence to security and the future of work. RAND.

Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A (2017) Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security. ACM, pp 506–519. https://doi.org/10.1145/3052973.3053009

Pariser E (2011) The filter bubble: how the new personalized web is changing what we read and how we think. Penguin, New York

Polya G (1945) How to solve it: a new aspect of mathematical method. Princeton University Press, Princeton

Pressler SJ (2016) Women with heart failure are disproportionately studied as compared with prevalence: a review of published studies from 2013. J Cardiovasc Nurs 31(1):84–88

Price R (2016) Microsoft is deleting its AI chatbot's incredibly racist tweets. Bus Insider. https://www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3. Accessed 17 Jun 2020

Raaijmakers S, Sappelli M, Kraaij W (2017) Investigating the interpretability of hidden layers in deep text mining. In: Proceedings of semantics 2017, Amsterdam

Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon JF, Breazeal C, Crandall JW, Christakis A, Couzin ID, Jackson MO, Jennings NR, Kamar E, Kloumann IM, Larochelle H, Lazer D, McElreath R, Mislove A, Parkes DC, Pentland AS, Roberts ME, Shariff A, Tenenbaum JB, Wellman M (2019) Machine behaviour. Nature 568(7753):477

Reilly P (2018) The impact of artificial intelligence on the HR function. In: IES perspectives on HR 2018, Institute of Employment Studies, UK, Member Paper 142

Rushkoff D (2010) Program or be programmed: ten commands for a digital age. Or Books

Rushkoff D (2019) Team human. Norton & Company Inc, New York

Russell S (2019) Human compatible: artificial intelligence and the problem of control. Penguin, New York

Russell S, Dewey D, Tegmark M (2015) Research priorities for robust and beneficial artificial intelligence. AI Mag 36(4):105. https://doi.org/10.1609/aimag.v36i4.2577

Sage A (2018) Waymo unveils self-driving taxi service in Arizona for paying customers *Reuters*. Bus News. https://www.reuters.com/article/us-waymo-selfdriving-focus-idUSKBN1O41M2. Accessed 17 Jun 2020

Santos JC, Tarrit K, Mirakhorli M (2017) A catalog of security architecture weaknesses. In: 2017 IEEE international conference on software architecture workshops (ICSAW). IEEE, pp 220–223

Schank R (2017). Ten questions for (and about AI). https://www.linkedin.com/pulse/ten-questions-ai-roger-schank/. Accessed 20 Mar 2019

Searle JR (1980) Minds, brains, and programs. Behav Brain Sci 3(3):417–424

Sesay A, Steffen J (2020) Wearables as augmentation means: conceptual definition, pathways, and research framework. In: Proceedings of the 53rd Hawaii international conference on system sciences

Sharkey N (2017) Why robots should not be delegated with the decision to kill. Connect Sci 29(2):1771–1786. https://doi.org/10.1080/09540091.2017.1310183

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

Simonite T (2017) Sorry, banning 'killer robots' just isn't practical. Wired. https://www.wired.com/story/sorry-banning-killer-robots-just-isnt-practical/. Accessed 22 Aug 2017

Sloman S, Fernbach P (2018) The knowledge illusion: why we never think alone. Penguin, New York

Smirnov A, Ponomarev A (2019) Decision support based on human-machine collective intelligence: Major challenges. In Internet of Things, Smart Spaces, and Next Generation Networks and Systems (pp. 113-124). Springer, Cham

Smit, S. K., Vries, A. D., Kleij, R., & van Vliet, P. J. (2016). Van predictive naar prescriptive policing: Verder dan vakjes voorspellen. TNO, The Hague

Streitfeld D (2018) Computer stories: A.I. is beginning to assist novelists. New York Times. https://www.nytimes.com/2018/10/18/technology/ai-is-beginning-to-assist-novelists.html

Surden H, Williams MA (2016) Technological opacity, predictability, and self-driving cars. Cardozo L. Rev 38:121

Sutton J, Harris CB, Keil PG, Barnier AJ (2010) The psychology of memory, extended cognition, and socially distributed remembering. Phenomenol Cognit Scinces 9(4):521–560

Thai J, Laurent-Brouty N, Bayen AM (2016) Negative externalities of GPS-enabled routing applications: a game theoretical approach. In: 2016 IEEE 19th international conference on intelligent transportation systems (ITSC), 595–601. https://doi.org/10.1109/ITSC.2016.7795614

Theiner G, Allen C, Goldstone RL (2010) Recognizing group cognition. Cognit Syst Res 11(4):378–395

van Diggelen J, Neerincx M, Peeters M, Schraagen JM (2018) Developing effective and resilient human-agent teamwork using team design patterns. IEEE intelligent systems, 34(2), 15-24

van Diggelen J, Barnhoorn JS, Peeters MMM, van Staal W, van Stolk M, van der Vecht B, van der Waa J, Schraagen JM (2019) Pluggable social artificial intelligence for enabling human-agent teaming. arXiv:1909.04492

van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, Heymann D, Burke DS (2014) A systematic review of barriers to data sharing in public health. BMC Public Health 14(1):1144

Surowiecki J (2005) The wisdom of crowds. Anchor Books, New York

van Wynsberghe A, Robbins S (2018) Critiquing the reasons for making artificial moral agents. Sci Eng Ethics. https://doi.org/10.1007/s11948-018-0030-8

Wallis KF (2014) Revisiting Francis Galton's forecasting competition. Stat Sci 29:420–424

Weise E (2017) Waze and other traffic dodging apps prompt cities to game the algorithms. USA Today. https://eu.usatoday.com/story/tech/news/2017/03/06/mapping-software-routing-waze-google-traffic-calming-algorithmsi/98588980/. Accessed 17 Jun 2020

Weizenbaum J (1966) ELIZA—a computer program for the study of natural language communication between man and machine. Commun ACM 9(1):36–45

Winter C (2017) "Killer robots": autonomous weapons pose moral dilemma. Deutsche Welle. https://www.dw.com/en/killer-robots-autonomous-weapons-pose-moral-dilemma/a-41342616. Accessed 17 Jun 2020

Wolfson S (2018) Amazon's Alexa recorded private conversation and sent it to random contact. Guardian. https://www.theguardian.com/technology/2018/may/24/amazon-alexa-recorded-conversation. Accessed 17 Jun 2020

Woods DD (2016) The risks of autonomy: Doyle's catch. J Cognit Eng Decis Mak 10(2):131–133

Woolley SC, Howard PN (2017) Computational propaganda worldwide: executive summary. In: Woolley S, Howard PN (eds) Working Paper, 2017. Project on Computational Propaganda, Oxford. https://comprop.oii.ox.ac.uk/

Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW (2010) Evidence for a collective intelligence factor in the performance of human groups. Science 330(6004):686–688

Yampolskiy RV, Spellchecker MS (2016) Artificial intelligence safety and cybersecurity: a timeline of AI failures. http://arxiv.org/abs/1610.07997 [Cs], October 25, 2016. http://arxiv.org/abs/1610.07997

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