AI & SOCIETY
https://doi.org/10.1007/s00146-022-01545-5

OPEN FORUM

AI and society: a virtue ethics approach

Mirko Farina[1] · **Petr Zhdanov**[1] · **Artur Karimov**[2] · **Andrea Lavazza**[3]

Received: 9 April 2022 / Accepted: 26 July 2022

Abstract
Advances in artificial intelligence and robotics stand to change many aspects of our lives, including our values. If trends continue as expected, many industries will undergo automation in the near future, calling into question whether we can still value the sense of identity and security our occupations once provided us with. Likewise, the advent of social robots driven by AI, appears to be shifting the meaning of numerous, long-standing values associated with interpersonal relationships, like friendship. Furthermore, powerful actors' and institutions' increasing reliance on AI to make decisions that may affect how people live their lives may have a significant impact on privacy while also raising issues about algorithmic transparency and human control. In this paper, building and expanding on previous works, we will look at how the deployment of Artificial Intelligence technology may lead to changes in identity, security, and other crucial values (such as friendship, fairness, and privacy). We will discuss what challenges we may face in the process, while critically reflecting on whether such changes may be desirable. Finally, drawing on a series of considerations underlying virtue ethics, we will formulate a set of preliminary suggestions, which—we hope—can be used to more carefully guide the future roll out of AI technologies for human flourishing; that is, for social and moral good.

Keywords AI technologies and applications · Virtue ethics · Human flourishing · Machine learning

1 Introduction

In the first half of the twentieth century, science fiction and Hollywood movies already envisaged an active role of intelligent machines in society (Bench-Capon et al. 2012). It all began, one may say, with Tin Man the famous character in the fictional Land of Oz created by American author L. Frank Baum, and continued with Maria—the humanoid robot that played in Metropolis—Fritz Lang's 1927 classic (Buchanan 2005). It is safe to say that by the 1940s, an army of computer scientists, mathematicians, and philosophers was actively trying to build intelligent machines and that the concept of artificial intelligence had already been culturally assimilated in people's minds (Brooks 1999). Vannevar Bush's visionary work (Bush 1945) is particularly instructive in this respect, as it anticipated many aspects of the modern information society. Bush's work was also inspirational for one of the fathers of Computer Science, Alan Turing. Alan Turing was a British polymath, who established the foundations of modern computing. Turing believed that computers could think and would eventually be able to possess human level intelligence. In particular, he suggested that human cognition can be understood as a mere computation or manipulation of symbols (Turing and Haugeland 1950). If we could map the operations of the brain formally (mathematically), then—on his view—we could also process them into a Turing machine (an abstract computational device whose operations are limited to reading and writing

✉ Mirko Farina
m.farina@innopolis.ru; mirko.farina@kcl.ac.uk
http://mirkofarina.weebly.com/

Petr Zhdanov
pe.zhdanov@innopolis.ru

Artur Karimov
arrkarimov@kpfu.ru

Andrea Lavazza
lavazza67@gmail.com
https://www.cui.org/andrea-lavazza/

[1] Faculty of Humanities and Social Sciences, Innopolis University, Universitetskaya St 1, Innopolis, Republic of Tatarstan 420500, Russian Federation

[2] Department of Social Philosophy, Kazan Federal University, Kremlevskaya St., 18, Kazan, Republic of Tatarstan 420008, Russian Federation

[3] Centro Universitario Internazionale, Via Antonio Garbasso 42, 52100 Arezzo, AR, Italy

Published online: 10 September 2022

symbols on a tape, or moving along the tape left or right); so, we could practically replicate human behaviour (and intelligence) in such devices (Copeland 2000).

Turing's groundbreaking idea faced two insurmountable challenges, though. Firstly, computers at that time could not store information, they could only execute commands. The capacity of storing information and retrieving it at will is (among other things) a crucial feature of human intelligence (Sternberg 1983). Secondly, computing was extremely expensive. In the early 1950s, the cost of leasing a computer was around USD 200,000 a month (approximately USD 2,306,622.41 in today's money, if we consider inflation). This meant that only very prestigious universities or big tech companies could afford it. Proofs of significance and further discoveries were needed to persuade funding agencies to finance projects on machine intelligence (Kline 2010).

A significant breakthrough came about during the early 1950s, with the development of the *Logic Theorist*. The Logic Theorist was a computer program designed to replicate the problem solving skills typically found in humans (Norvig and Intelligence 2002). This program, arguably one of the first artificial intelligence programs ever designed, was presented at *the Dartmouth Summer Research Project on Artificial Intelligence* (DSRPAI) in 1956. This conference—a collaborative effort who brought together top researchers from various academic fields, was jointly hosted by John McCarthy and Marvin Minsky. During this conference the term '*artificial intelligence*' was coined (McCorduck and Cfe 2004). Despite the Dartmouth Summer Research Project on Artificial Intelligence fell short of its very ambitious goals (McCarthy et al. 1956), it set the agenda by catalyzing the next 20 years of AI research.

As machines became faster, cheaper, and more accessible, the field literally boomed. From 1957 to 1974, research on AI flourished. Newell and Simon's *General Problem Solver* (Newell et al. 1959), Rosenblatt's *PERCEPTRON* (Rosenblatt 1960) and Weizenbaum's *ELIZA* (Weizenbaum 1976) [to mentioned only a few] showed realistic potentials to replicate problem solving skills and learning abilities into artificial systems, sparking widespread enthusiasm about the future prospects of AI. In the late 1960s, Minsky famously noted that "from three to eight years we will have a machine with the general intelligence of an average human being". Building and expanding on this seminal work, Newell and Simon (1972) formulated the physical symbol system hypothesis, which laid down the groundwork for the development of the computational theory of mind (Fodor 1975); a very influential paradigm in philosophy, which attempted to explain human cognition in terms of computation performed on internal (neurally localizable) representations.

Nevertheless, while much of the theoretical apparatus for the development of intelligent systems had been developed, there was still a long way to go before the end goals of abstract thinking and learning could be successfully implemented in artificially constructed machines. Devastating critiques quickly followed those very early successes, pointing out the insurmountable difficulties (both technical or mathematical—Minsky and Papert 1969 and philosophical—Searle 1982; Dreyfus 1976) underlying the development of a strong AI (Dreyfus and Hubert 1992). As hopes dwindled so did the funding, and research on AI came to a halt for about 10 years.[1]

It was only in the late 1980s and early 1990s, with the advent of "deep learning" techniques and back-propagation training (Hopfield 1982; Rumelhart et al. 1986), that researchers found a way out of the impasse above-mentioned, proving that computers could actually learn from experience and might ultimately be able to attain human like intelligence (McClelland et al. 1986; Rumelhart et al. 1994). For example, in 1999 LeCun et al. (1999) pioneered the concept of convolutional neural networks.

During the 1990s and early 2000s, many landmark goals of artificial intelligence were achieved. For instance, in 1997, Gary Kasparov (chess grand-master and world champion) was defeated by IBM's Deep Blue. In the same year, a speech recognition software, developed by Dragon Systems, was firstly implemented on Windows, opening the way to the development of modern virtual assistants (such as *Siri, Alexa, Alisa* etc.) capable of interacting and communicating with humans using speech synthesis. Similarly, artificial machines (such as *Kismet*) became capable of mapping and recognising human emotions through facial analysis (Breazeal 2002).

We now live in the age of "big data"; an age in which we have the capacity to collect huge amounts of information and process it with artificial machines in ways that are far more effective than a person (or a group of people) could ever do. In this context, Krizhevsky et al. (2012), Dean et al. (2012) further demonstrated the great power of deep learning techniques for several industries and various fields (ranging from engineering and banking, to marketing, and even entertainment); see Le et al. (2020) for a helpful review. In the long term, the goal of researchers working on AI is to produce systems and artificial machines that will be capable of exceeding and surpassing human cognitive abilities in virtually all tasks or fields. This possibility, thanks to the above-mentioned progresses and heavy investments by governments and funding agencies worldwide, seems within reach, or at least it no longer looks like an unattainable mirage.

This possibility also promises to bring about a fourth industrial revolution (Philbeck and Davis 2018) and with it

[1] https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/ (Last Accessed September 2022).

large-scale changes to the functioning of our society, potentially leading to undesirable societal outcomes (e.g., unusable skills and job losses, social insecurities, further centralization of power and money, concerns over privacy, just to mention a few). For this reason, a number of governments (e.g., Chinese,[2] American,[3] and Russian,[4] to mention a few) as well as institutions (such as the EU[5]) advised by leading researchers (Floridi et al. 2018; Floridi 2019; Floridi and Cowls 2022; Cowls et al. 2019) already started evaluating the grand challenges that AI may pose to our societies (Yang et al. 2018; Taddeo and Floridi 2018) as well as reflecting on the opportunities to improve the well-being of their citizens (Vallor 2016, 2017; Walsh et al. 2019), by promoting and endorsing the application and design of AI for Social Good (Floridi 2020; Smuha 2019; Delcker 2018) and Moral Enhancement (Savulescu 2009; Savulescu and Maslen 2015; Clarke et al. 2021).

Naturally, as noted above, a lot of works have been produced in the area and many important intellectuals analysed these problems before us. In this paper,building and expanding on those earlier works, we reflect on the possibility and desirability of applying new AI technologies in society. More precisely, in Sect. 2, we discuss a methodology—recently applied in software engineering and based on Machine Learning—called Puzzle Driven Development (*PDD*), which (and more on this below) aims to automatically prioritize tasks allocation for software engineers. We review the pros and cons of its widespread application in workplaces (such as IT companies) and on the job market. In Sect. 3, we extend the reach and scope of our analysis to other domains (such as *social robotics, criminal justice, and medicine and healthcare*) with the aim of better comprehending the broader potential impact of AI technologies in our societies. In doing so, we critically reflect on the opportunities (*promises*) as well as on the challenges (*dangers*) that the adoption of such technologies may bring about. In Sect. 4, by focusing on a series of important considerations underlying virtue ethics, we assess the desirability of this process (AI revolution), while developing a set of preliminary recommendations or suggestions, which—we hope—can be used to more carefully guide the future implementation of AI technologies for human flourishing; that is for social and moral good. Finally, we conclude (Sect. 5), by discussing possible future research directions.

2 AI in the workplace: puzzle driven development

Writing software is a complex process that typically requires several programmers working collaboratively with the same code base (Wasserman 1996). Version Control Systems (VCS) are used by most projects to maintain a code base (e.g., Git). So called "feature branches" provided by VCS help developers isolating their changes and subsequently testing and integrating them into the source code trunk (Spinellis 2005). However, when a development team is large and the intensity of the changes is significant, developers often face the so-called problem of stale branches (branches of the repository that have become too large and haven't been dealt with by anyone in a very long time).[6]

To prevent such situations from happening and, hence, to streamline the development process, some teams adopt early merges of incomplete changes. In other words, they force programmers to leave TODO markers (known as puzzles) in any unresolved branch (Storey et al. 2008). These puzzles are then converted (typically by the project manager) to new tasks when the branch gets merged into the trunk. Even though there are a few existing tools developed to automate this process, none of them—to date—can prioritize and properly assign the tasks being created (Guo et al. 2021). Because of this significant shortcoming that causes tasks to quickly flood the backlog, such tools haven't been used in large and complex projects (Schwaber 1997).

Recently, however, a few researchers proposed a new development methodology called Puzzle Driven Development.[7] This technique (Bugayenko et al. 2022a, b) delegates the responsibility of task decomposition to its performers (the developers). In doing so, it eliminates the role of the project manager in task decomposition. In other words, this technique manages the programmer and their tasks at the same time, by requiring the maximum amount of work that can be done by any developer on any given task. This means that the developer focuses on delivering the task for which she has been hired by contributing the maximum to it and with the least amount of resistance (like blockers, unclear descriptions, dependencies on other tasks). The developer then creates more tasks by adding puzzles to positions in the codebase where the rest of the work needs to be done. This implies that the initial task can be moved to a "done" or "completed" state and its children tasks (created using

[2] http://www.gov.cn/zhengce/content/2017-07/20/content5211996.html (Last Accessed September 2022).

[3] https://www.ai.gov/ (Last Accessed September 2022).

[4] http://en.kremlin.ru/events/president/news/57425 (Last Accessed September 2022).

[5] https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html (Last Accessed September 2022).

[6] An example at: https://github.com/git/git/commit/e194cd1e0e08611462eb9c5a731a7a3d797f9252 (Last Accessed September 2022).

[7] https://www.yegor256.com/2018/03/21/zerocracy-announcement.html (Last Accessed September 2022); https://patents.google.com/patent/US20120023476A1/en (Last Accessed September 2022).

puzzles) do not have a hard dependency on the parent task. In essence, these newly created tasks are not blocked by the initial task.

Complexity and volume are often mentioned as two of the major factors attributed to incomplete, delayed and/or failed tasks/projects in software development (Mitchell 1990). Smaller and simple tasks with shorter estimated delivery times, on the other hand, have a higher completion and success rate (Lalsing et al. 2012). Crucially, effective task management is of paramount importance in software engineering, especially for large organizations with hundreds of thousands of developers and significant amounts of backlogs. Projects managers, though, are often too busy to deal with all the tasks of a project because they often manage more than one project at the same time. The PDD methodology promises to enable the continuous free flow of work, while limiting dependencies between tasks or blockers, which often lead to delays in task delivery and subsequently, to significant delays in project delivery. In brief, with the use of Machine Learning (ML) it is now virtually possible to automatically prioritize tasks and adopt smarter forms of management that help increasing the effectiveness as well as the productivity of software development (Ciancarini et al. 2021a, b).

In truth, business companies have long used AI-based solutions to automate routine tasks in operations and logistics (Raisch and Krakowski 2021). However, with recent advances in computational power, the exponential increase in data, and new machine-learning techniques—such as the one we discussed above—corporations are now able to use AI-based solutions also for important managerial tasks (Brynjolfsson and Mcafee 2017)

It is therefore not surprising that big IT companies are heavily investing in these techniques and methodologies, foreseeing the unique potentialities of Machine Learning for streamlining workflow, efficiency, as well as for maximising profits. Admittedly, the adoption of Machine Learning, if properly implemented, could also benefit the workers. For example, it could contribute to significantly reduce stress, thereby allowing developers to achieve a more balanced lifestyle and to live fuller lives. In this sense, we could positively assess such developments and consider them as illustrative, paradigmatic examples of the rising tide of artificial intelligence technology to business automation, which also currently involves other fields (Wright and Schultz 2018) (ranging from robotics and entertainment to medicine and sensors).

Nevertheless, despite these extraordinary applications and the promise to enhance workers' well being, it is still crucially important to adopt a critical stance and encourage some criticism towards the widespread implementation of such techniques and technologies in the workplace. Thus, it is of paramount importance to reflect on the potential nefarious ethical and sociological implications that business automation may have for future professions, workplaces, and even civil rights. Next, we turn to analyse some of these effects or implications.

A great number of researchers already pointed out the potential negative impact of automation on the future of jobs (Brynjolfsson and McAfee 2014; Susskind and Susskind 2015). Some of them argued that this digital revolution is going to create massive job losses on an unprecedented scale (see Wajcman 2017 for an helpful review). Others (e.g., Ford 2017) claimed that due to information technology not only low-skill workers but also highly-skilled professionals will be at risk of being displaced by machines. Most of these researchers therefore share the view that automation is likely to produce significant labor disruptions (e.g., Arntz et al. 2017). For example, Frey and Osborne (2017) predicted that automation could replace as much as 50% of today's jobs in as little as 10 years.

Not everyone has looked at this process with negative eyes, though. For example, a few researchers envisaged favorable labor supply adjustments following this AI revolution (Abeliansky and Prettner 2017; Kurzweil 2005).

Some researchers also took a (pragmatic) middle ground position (Daugherty and Wilson 2018; Davenport and Kirby 2016). They emphasised the many limitations that machines still have and anticipated the advent of an era where, rather than being adversaries, humans and artificial systems will coalesce, combining their complementary strengths to enable faster learning and the development of enhanced capabilities (Markoff 2016).

Consistent with these recommendations, many IT companies started pursuing the implementation of an augmentation strategy, rather than one involving mere automation. For example, Satya Nadella, CEO of Microsoft, stated that the firm will "build intelligence that augments human abilities and experiences. Ultimately, it's not going to be about human vs. machine".[8] On a similar vein, IBM recently asserted that "the purpose of AI and cognitive systems developed and applied by the IBM company is to augment human intelligence".[9] Yet, significant worries remain concerning the fact that AI technology (such as machine learning), while improving the economy, streamlining efficiency, and maximising profits may also exacerbate and radicalise societal inequalities, by—for instance—reducing employment and wages, especially for the working and the middle

[8] https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html (Last Accessed September 2022).

[9] https://www.ibm.com/blogs/think/2017/01/ibm-cognitive-principles/ (Last Accessed September 2022).

class (Acemoglu and Restrepo 2020), thereby leading us into a dystopian future (Bostrom 2017).

However, returning to the case study in software engineering (PDD) we presented above, we can say that what is truly original and perhaps more interesting about such a case is not necessarily the fact that it will lead to increase automation; rather, that it can be taken to show how automation may become an essential requirement for the implementation and functioning of the process itself. In other words, the software/the program/the technique won't merely be a tool in the hands of IT companies to increase efficiency, decrease costs, or speed up the development process. In brief, something we—as humans—may choose to implement or refuse to based on a series of partial and subjective objectives. The PDD case is potentially illustrative of an economy and of a society that, once built, wouldn't be able to give up the software or the programs on which it is be based. This is because doing so, in virtually every domain (from communication to public health, and finance), will determine a collapse of the society itself. This means that it is not just the realisation of the infrastructure that leads to automation that matters in this case, but also the processes managing the update and repair of the software/programs as well as those that determine their functioning that become crucial. This is because such processes are mandatory to keep the system, hence the society based on them, alive; that is, up and running. The pressure then to speed up and streamline everything that relates to the software or the program (even the capacity to fend off cyber-attacks) becomes an imperative, a must do to ensure the system's survival. In this sense, the PDD (and similar cases) may constitute obligations that must be carried out, in the most efficient way possible, even if they may cause significant job losses. This seems to be a very peculiar characteristic of an economy that is based on information technology, which is—however—also deeply welded on a significant component of AI, to the extent that—one may argue—makes it utterly dependent upon it.

The possible dependence of future societies on increasingly powerful AI systems raises another significant worry. The concern one may raise with respect to this point is that AI technologies are not just any piece of technology—they are pieces of technology (that, at least in some cases, such as the one we discussed above) are meant to fully replace humans in their capacity of decision-making. There are several important ethical questions that one may ask in this respect. For instance: (i) what risks are involved when AI replaces humans in their capacity of decision-making, and (ii) how do we best avoid such risks? or (iii) is ethically desirable or permissible to delegate important decisions to AI systems? Provided that AI is more efficient than human intelligence for certain tasks, it must be noted that decision-making in humans is a complex process, which is often based on moral values and ethical considerations of things such as respect, dignity, non-discrimination, etc., that contemporary AI systems lack. While these issues may not be directly relevant in cases where AI is used for human-machine interaction, they become extremely significant in cases where AI is deployed for human-human interactions (such as in some of the cases we will discuss below).

3 AI in society: promises and dangers

In the nearest future, AI as a technology will likely change not only the way we operate at work, but also the way we live our lives. In general, one may venture to say that the prosperity of future societies may well depend on their abilities to adapt to the transformational changes that the AI revolution will bring about. In this section we focus specifically on three domains and on the analysis of how AI may influence specific human values related to them: (a) social robotics (friendship), (b) criminal justice (fairness), and (c) medicine and healthcare (privacy).

3.1 Social robotics

In the last decade, there have been significant efforts to study the possible uses of social robots (Johal 2020; Heerink et al. 2016). These are robots built in a humanoid or animal-like form that are developed for natural, affective, and quasi-interpersonal human–robot interactions (Breazeal et al. 2016). For such interactions to be akin to the interpersonal interactions observed among humans, social robots need to be equipped with human-like emotions, which are usually developed using methods of artificial intelligence, known as artificial emotional intelligence (AEI). Such methods typically allow or assist social robots in performing the following tasks: *i. emotion recognition, ii. emotion generation, and iii. emotion augmentation* (Schuller and Schuller 2018). While the first two of these tasks are actively studied and even partially implemented in social robotics; the task of emotion augmentation has remained—at the time of writing at least—partly unexplored. However, we must note that even the simplest task; that of emotion recognition may—to date—represent a challenge, especially when considering 'microexpressions' on human faces, for examples signals of power (posture, volume of voice, etc.; see Krakovsky [2018]). In addition, given the multi-modal nature characterising much of human communication, it is absolutely critical for AI (if it wants to be successful, at least) to correctly recognize and understand human psychology. In this context it is worth noting that consultative AI, the sort of AI that is extensively used in information assistance services, can rarely communicate information to calm down a person if the person shows signs of irritation or stress. Similarly, most of the social robots available to date are able

to perform specific roles (e.g., involving caring); however, all such roles are typically performed on a structured and predictable scenario where psychological variables are not taken into account. Even though social robots may offer certain benefits- for example- they could alleviate stress or other negative emotions or even contribute to develop empathy in children, as recently shown by Alemi et al. (2016); they may not be able to repeat some of the same positive effects in adults, unless adult humans consider such robots smart enough to sympathize and engage with them (in other words, unless such robots demonstrate the capacity to recognize emotions and to act in accordance with social norms). However, as interest in social robots will likely increase, it may be worth reflecting on the possibility of using social robots as means to cultivate friendship. Many researchers (Gabriel 2020; Montes and Goertzel 2019; Renda 2019) already speculated about the possibility to create robot-friends (Marti 2010). For example, Promobot recently introduced a new humanoid robot capable of gesticulating and communicating realistically with people, pretty much like a friend.[10]

With high hopes, however, also came strong criticism. One concern, raised by Danaher (2019), is that robots will only -at best- be able to imitate the emotional reactions and behavioral patterns that humans associate with friendship. Even though interaction-oriented robots may seem to become more human-like and with the introduction of AI they will able to communicate in a more personalized manner, their social abilities will still lag behind those of a human being. As a consequence, one may conclude, that the human–robot experience will still significantly differ from the kind of friendship observed among humans.

In this context, we can raise a second concern, which unlike the one we presented above, has not been widely discussed in the literature and as such is quite original. The concern is not about the inability of social robots to mimic behavioural patterns observed in friendship among humans but about their incapacity to mimic the styles of friendship observed among real friends. Consider a lonely elderly or a person living in a remote (isolated) place (e.g., in the mountains), a humanoid robot friend might certainly be a desirable solution for such a person (certainly a better solution than its alternative; that is, no friends at all); however, the ethically sensitive aspect that we think needs to be discussed concerns the friendship style that the robot must be trained/pre-programmed to pursue or exercise.

Should it be "confirmatory and caring"; i.e., should it always indulge and support the human? Or should it be also "corrective"; that is, by virtue of its "neutral" and logical competence (the robot is still a computer) must it try to convince the human to modify or to improve certain attitudes, behaviours, and/or lifestyles when logic demands it? Should the robot perhaps be "adversative"; i.e., possess its own point of view (which one, in case?) and support it in the interaction with the human even when it leads to a clash? In brief, the point we are after here is that the human friend does a bit of all of the above and probably does even more. For example, the human friend is capable not only of switching its behaviour/stance (from, say, confirmatory and caring, to neutral/corrective, or even adversative) but also to adjusting it with respect to changing environmental circumstances and/or psychological scenarios. The robotic friend, on the contrary, is neither capable of changing its stance (to switch modality, unless reprogrammed) nor is able to adjust its behaviour in relation to changing conditions (be them environmental or psychological). It simply performs the job for which it has been trained or pre-programmed. In this context, it may well happen that the robotic friend that panders and indulges and supports on all conditions (no matters what) its human friend may be preferred over, say, an overly critical human being. This situation, however, poses significant ethical issues, as it may gradually contribute to a loss of important societal values (such as honesty, respect, integrity and other normative ones) that appear to characterise real friendship but that are also of paramount importance for proper and harmonious group relations.

3.2 Criminal justice

Another domain where the AI is currently being implemented is criminal justice. 'Risk assessment tools' are increasingly used in courts for dealing with a number of tasks such as: (i) assisting judicial decision-making, (ii) initial sentencing, (iii) probation and parole granting, and (iv) post-release monitoring (Weidong 2020; Wang 2020). The introduction of AI in criminal justice serves mainly the purpose of automation, pursued in the name of objectiveness and neutrality. As automation is based on datasets, statistical models or inferences, it is believed that AI could elaborate more objective data, which would allow it to render fairer judgements.

In this respect there are a series of important ethical considerations to make. Firstly, how reliable are those datasets and statistical models? The crimes recorded at police departments are not necessarily all the crimes that happen, as the criminality rate may depend (quite heavily) on policing strategy, and also on relations between the local community and the police (Isaac 2017). Thus, one of the most critical and crucial aspects underlying the implementation of AI technologies in this domain involves the reliability of the algorithmic computations that are fed to the AI system and their potentially discriminative nature. If such datasets and algorithms or statistics are not proper and sound, they

[10] https://tadviser.com/index.php/Product:PromobotRobo-C (Last Accessed September 2022).

may well cause extensive false positives, especially towards ethnic minorities (Douglas et al. 2017). What may potentially happen is—for instance—that stringent post-release policing in certain areas with already high criminality rate may result in mass incarcerations, while in other areas less stringent policing policies may just cause the AI system to release pressure on criminals, allowing them to act more or less freely. On these grounds, one could argue that the implementation of automation and AI in criminal justice may actually threaten (rather than ameliorating) the functioning of the criminal justice system.

Secondly, one could also raise another worry, related to the one discussed above, which is about the application of AI in the determination of risk for bail and parole; or in the usage of AI for sentencing. Surely, a positive bias towards automation could be defended in all these cases. If we think about how wrong human inferences can be (Kahneman 2011) or about how a judges' eating habits may affect the harshness of punishment (Glöckner 2016; Danziger et al. 2011), the idea of automating the process may quickly arise as the fairest as well as the most efficient prospect available. However, a judgement made through an algorithm is not always better than a judgment performed by a human being, as we already explained above, especially because—in the end—humans are strongly biased creatures, who invariably prefer evaluations made by other humans (especially when it comes to punishment) (McKay 2020). So, any decision (however fair) made by a machine (however intelligent) may never be fully accepted by a human being and this may lead to potentially significant social tensions. All in all then, there seems to be serious issues underlying the implementation of AI technology in the criminal justice systems, which may threaten both ethical and social values.

3.3 Medicine and healthcare

In the area of medicine the primary role of AI is to analyze the relationships between prevention or treatment techniques of a wide array of medical conditions, and/or to improve medical diagnosis [Ahuja (2019) and Sand et al. (2021)]. There is no doubt that the intelligent clustering of medical data can assist doctors in many ways. For example, computational psychiatry (Montague et al. 2012) uses computational tools (such as machine-learning) to process the patient's clinical data in order to improve disease classification, predict treatment outcomes and -on these bases- refine the choice of drugs to be administered (Wiese and Friston 2021).

The ethical aspect underlying the application of AI in this new branch of medicine is related to the possible "depersonalization" of treatment, which is based on the digital processing of quantitative parameters, without considering the phenomenological aspects that are related to the physiological ones. Although a few researchers recently argued that there are potential significant benefits in this process (Palmer and Schwan 2021), especially related to the mitigation of shame-induced barriers to medical care, others [such as Alvarado (2021)] also convincingly showed that even if deep neural networks trained by AI may be better than human doctors with respect to accuracy (at least in certain domains; e.g., radiology), may be less costly, and even possess more predictive power; we would not have sufficient reason to trust them in the same way as we trust a human counterpart. In other words, even if AI and its computational techniques may be able to trigger otherwise unattainable advancements in the way we understand the biochemical balances of our brain, this does not necessarily mean that the individual will trust them and that humans will enjoy an equally satisfactory experience 'from the phenomenological perspective', at least (Wiese and Friston 2021).

Yet, one may still argue that the pattern-recognition technology typically used in machine learning, combined with the opportunity of testing on 'virtual cohorts', could provide very powerful opportunities for the development of better treatment in several branches of medicine (Altman 2015). This might be true. However, in all the cases we discuss there is a significant amount of personal data to dealt with (the patients' data). Hence, it seems crucial to determine the boundaries of privacy, so that the data would allow training the machine and carrying out its assessment on a condition of anonymity and in accordance with ethical principles and values. It is worth noting that many researchers already started investigating the ways in which Blockchain technology could be applied in healthcare, for instance (Xia et al. 2017). Blockchain technology, in tying the possibility of accessing personal data to the person's biometrics, could well be used to place the patients at the center of the medical ecosystem, to increase their security, privacy, and even the interoperability of their health data (Megha et al. 2021).

However, how secure the personal information recorded on that blockchain could ever be? Otherwise stated, how a blockchain system is coded and funded, to whom it ultimately responds, and under which legal framework it works would have fundamental intentional and unintentional ethical consequences on the way in which the person's privacy is preserved (Megha et al. 2021). This area, at the time of writing, is simply too murky to make any sensible or definite ethical recommendation on the topic.

As a final note to this discussion, one could notice that in cases of human-machine interaction (Farina and Lavazza 2021a, b; 2022a, b) the risks involved in the deployment of AI systems may be rather limited (related—at worst—to job losses), while they might be much more significant if the AI system is meant to replace the human in human interactions. This is because artificial machines are not autonomous moral agents, they don't have the ability—at the time

of writing at least—to intend an action or to autonomously choose an intentional action; so, they lack the deliberation required to make responsible decisions. Thus, relevant ethical considerations should be built in such systems by the developers themselves, possibly as behavioural patterns, as it seems clear that ethical insights are desperately needed for such systems.

4 Designing AI for moral and social good: a virtuous ethics perspective

Having critically assessed the opportunities and the challenges underlying the implementation of AI technologies in the wider society, we can now briefly reflect on how to use such technologies for social and moral good (Cath et al. 2018).

Although—as Jones (2013) brilliantly noticed—there is probably a little Ned Ludd in us all, and even though each of us could certainly sympathize with those workers in England who 'toiled two centuries ago in "dark Satanic mills," as poetically described by Blake and Bloom (1982),[11] and felt driven to smash the machines that were replacing them'; it is practically impossible to think or even imagine that we could now meaningfully stop the revolution pushed forward through the development of AI technologies.

Wiener (1988) famously made the ethical argument that humans should be liberated from work that machines can do better. Yet, there is a widespread anxiety (call it paranoia, if you like) in society, visualised in many movies (ranging from Chaplin's *Modern Times* and Kubrick's *2001: A Space Odyssey* to the Wachowskis' *Matrix trilogy* and Cameron's *Terminator*), with respect to a future where machines will raise, revolt against us, and become our masters. This kind of understanding is also partly echoed in the philosophical writings of Frisch (1959) and Arendt (1950).

In truth, the ethics of AI has already been the subject of in-depth analyses that have explored many of the most significant aspects involved in it (Taddeo and Floridi 2018; Floridi et al. 2018; Floridi 2019; Cath et al. 2018; Cowls et al. 2019; Savulescu 2009; Savulescu and Maslen 2015).

Roughly speaking, we can say that there are three main directions or dimensions inspiring and characterising such ethical analyses (Li 2021).

1. *Consequentialist Ethics.* Under this framework an agent is thought to be (or become) ethical, if and only if—in weighing the consequences of each possible choice—she chooses the option that has the best aggregate consequences.
2. *Deontological Ethics.* Under this understanding an agent is believed to be (or become) ethical, if and only if she is conscious of her obligations and duties and—consequently—if she acts in accordance to established social norms.
3. *Virtue Ethics.* Under this framework an agent is (or becomes) ethical, if and only if she displays virtues (such as courage, justice, generosity etc.) and therefore acts according to exemplary moral values, so as to be perceived favourably by others.

Under any of these three frameworks, the need to maintain and preserve human self-determination (especially in relation to AI-governed systems) has been strongly emphasized (Coeckelbergh 2020; Spiekermann et al. 2022; Trappl, 2015). This is particularly important because we are entering an era when automation is being extended to many areas and domains, as we have seen above. Directly linked to the issue of self-determination is the issue of agency (the ability of human beings to act consciously and freely), which has also captured the attention of many researchers across the above-mentioned dimensions (Johnson and Verdicchio 2019). Agency is one of the properties that make individuals persons and allow them to have full moral status; hence, it is a property that is of paramount importance to attribute meaning to people's existence. This is why it has been noted that an ever-increasing extension of decision-making and control of complex processes by artificial intelligence systems may cause a restriction of human agency, albeit with a gain in efficiency (Hallamaa and Kalliokoski 2020).

Yet, it must be also noted that are more ethical issues that have received increased attention within those frameworks we mentioned above. These are those concerned with the possibility of explaining and making transparent the criteria on the basis of which an AI system can make decisions. For example, total human control of automation is obviously not a desirable outcome, as it seems to contradict the purpose for which AI is employed in the first place, which is—mainly—to do what a human being or even an organization of humans cannot do as effectively, efficiently, and quickly. Nevertheless, some form of control in terms of responsibility for the consequences of the decisions taken by an AI system seems to be necessary (de Fine Licht and de Fine Licht 2020; Felzmann et al. 2020). Such checks must be implemented in certain domains, for instance, in the medical applications of AI—mainly when there is the possibility of an inauspicious outcome. Is it an outcome that no doctor could have avoided in any case, or is it a case where the system has 'made a mistake'? And, consequently, who should bear the costs for the mistake?

[11] https://physicstoday.scitation.org/doi/10.1063/1.2731975 (Last Accessed September 2022).

The concern for transparency, i.e., understanding the ways in which the AI used reaches its decision, seems to be of paramount importance also with respect to the applications of AI technologies to the criminal justice system. Crucial points in this context involve reflections on: (i) why certain areas are deemed more in need of close police control than another?; or (ii) why an offender is given a harsher sentence than another offender, who has committed the same type of crime? These points, of course, become all the more relevant as the more invisible the artificial intelligence systems become. In truth, in this case it is not just the transparency of the automation process that seem to pose significant ethical issues but also the fact that some functions in those systems might be automated and hence be beyond human control.

In this sense, as discussed in Sect. 1 above, not only have many scholars and representatives of civil society voiced concerns and suggested forms of protection and regulation of people and society, but also some states and interstate organizations have established specific guidelines for the development of an ethical AI (Floridi et al. 2018). For example, a US report suggested specific policy responses to "amplify the best and temper the worst impacts" of AI and automation. A EU policy brief demanded for "intrinsically European and humanistic values" to ground "rules, governing in particular liability and ethics" of robotics and AI. A UK report focused on the importance of examining "the social, ethical, and legal implications of recent and potential developments in AI" and developing "socially beneficial AI systems" (Cath et al. 2018). On a similar vein, the code of Ethics developed by the Russian Federation[12] emphasized the need for AI systems to:

1. protect human interests and rights,
2. promote the responsible usage of AI technology in society,
3. pursue the implementation of social good, while
4. preserving maximal transparency and truthfulness about their capabilities and risks.

In an attempt to further these important considerations we may propose another reflection here. The starting point of which may be the idea that the AI seems to improve the effectiveness as well as the efficiency of many processes that humans strive to complete. In this sense, we should admit that AI can bring about positive innovations, precisely because of its enormous learning and processing capacity. So, one could suggest that a good society with a strong presence of AI could be one in which automation would not distort the natural characteristics of humans (although we cannot exclude that in the future human beings will undergo an evolution due to the interactions with digital devices). However, it must also be noted that humans—throughout their evolutionary history—have acquired some basic features in ancestral environments (such as the African savannah), which they continue to need (among others, one could mention: i. sociability and relationships, ii. the ability to frequent natural and not just artificial environments, and iii. various other activities (including physical efforts oriented to a relevant, concrete and visible purpose). These, we argue, are important conditions needed to be fulfilled to avoid psychic and existential problems that are more frequent in advanced societies and that are independent of the level of physical security, education, income, and wealth. Human beings are proactive creatures, who deeply fear loneliness, boredom, and feelings of worthlessness. In general, they like to be held accountable for their actions. The positive contribution of artificial intelligence to a good and fair society may then come from its integration with this feature of human beings, if pursued to enhance the needs of humans themselves (not to vex or mortify them). In other words, efficiency, automation, and optimization—even if carried out with the best intentions and with the idea of developing a more prosperous society in line with economic criteria and/or values, should always be balanced against the backdrop on what ultimately makes us 'phenomenologically' and 'quintessentially' humans. Otherwise stated, it should be directed towards preserving sociality, increasing happiness, and realizing human potential.

To this extent, we thus emphasise the need to focus on the application of AI technologies for human flourishing and well-being. In truth, a number of researchers, including (Constantinescu et al. 2021; Stahl 2021; Wallach and Vallor 2020) recently advocated such a view [which we may call virtuous ethics] and—as a consequence—a set of guiding principles (including beneficence, non-maleficence, autonomy, and justice) has emerged to guide and direct such implementation (Jobin et al. 2019).

A virtue-based approach typically focuses on certain character traits and dispositions, which are acquired through practice and habituation, rather than on rules and regulations. So, virtue ethics is said to be agent-centered rather than action-centered, it is about the right sort of feelings and motivations, rather than about the right sort of actions.

Consider next Shannon Vallor's work as a paradigmatic illustration of this approach (Vallor 2016, 2017). Vallor—in drawing from Aristotelian, Confucian, and Buddhist virtue traditions as well as from earlier works on virtue ethics conducted by Western philosophers (such as Nussbaum 1999; MacIntyre 1981)- proposes to cultivate a kind of moral character that expresses technomoral virtues. Such virtues—on her view—will allow us to live a good life in a future in

[12] http://publication.pravo.gov.ru/File/GetFile/0001202004240030?type=pdf (Last Accessed September 2022).

which technological powers become "embedded in co-evolving social practices, values, and institutions" (p. 5).

The technomoral virtues that can help flourish and thrive in the world, according to Vallor, include: (i) honesty, (ii) self-control, (iii) humility, (iv) justice, (v) courage, (vi) empathy, (vii) care, (viii) civility, (ix) flexibility, (x) perspective, (xi) magnanimity, and (xii) wisdom. Vallor maintains that these virtues will likely evolve in future technosocial contexts and that—hence—her taxonomy should not be considered as exhaustive.

Whether we should consider Vallor's taxonomy as an exhaustive one is beyond the scope of this paper; what is important in the economy of our manuscript is rather how Vallor thinks that such technomoral virtues can be acquired. The foundations of virtuous character, for Vallor, are mastered by promoting a "relational understanding of moral obligations, reflective self-examination of moral progress, and intentional self-direction of moral development" (Barrera 2020, p.129). In other words, one's moral wisdom is only completed through habitual moral attention to the salient features of specific situations coupled with the appropriate extension of moral concern to 'the right beings, at the right time, to the right degree, and in the right manner' (Vallor 2016, p. 117). In short, Vallor appeals to virtue ethics to offer us a strategy for self-cultivation that empowers us to ultimately develop "technomoral wisdom", a general condition of integrated, and unified moral expertise that genuinely expresses or synthesises all other technomoral virtues and does so towards the realization of the highest human good.

To understand and better appreciate the many benefits of adopting this virtuous approach to the applications of AI technologies in society, we next briefly review how it may perform against alternative accounts [consequentialist or deontological] (Bauer 2020; Anderson and Anderson 2011) in one of the case studies [involving social robots] that we discussed and analysed earlier on.

In considering the possibility of the implementation of a virtuous ethics approach in the field of social robotics, we should first distinguish between what we may call robust virtuous robotics and weak virtuous robotics. The robust approach to social robotics claims that it is theoretically and practically attractive to implement virtues in robots themselves, while its weak counterpart asserts that it is not needed that social robots behave in a virtuous way, it is simply enough that they serve for pedagogical purposes; that is, help humans cultivate virtues like compassion, honesty, and generosity (Peeters and Haselager 2021). The robust virtuous robotics has received some criticism (Constantinescu and Crisp 2022).

The crux of the argument against such an approach revolves around the idea that current or near-future social robotic systems cannot genuinely perform virtuously, nor have the right motivations to do so. In addition, it has been argued that such robots cannot acquire the phronesis needed for taking into account the right circumstances; hence, that they are unable to make appropriate moral decisions.Yet, we do not think that a weak virtuous approach in robot design is less preferable than alternative approaches (be them consequentialist or deontological)—quite the opposite. We believe that it is often undesirable to frame the relationship between humans and social robots through the language of duties and consequences.

Consider the following example as an illustration of this claim. Imagine that someone wants to simulate rape with a social robot (think about the famous "Westworld" tv series). What would be wrong with such an action from a deontological point of view? Do we have any moral duties towards the robot? Do standard moral obligations or universal precepts (such as do not rape etc) that normally apply to human relations also apply to our relation with the robot? Hardly so. Similarly, what would entail to assess such an action (simulating rape on a robot) from a consequentialist standpoint. Will there be bad consequences for anyone? Not really, none would suffer and nothing would really change for humans (although probably both deontologism and consequentialism advocates could find some arguments against robot rape). Yet, from the standpoint of virtuous ethics, we would immediately know that it wouldn't be virtuous to imitate rape using sex robots and that would intuitively go against basic exemplary traits of character that are valued by society and individuals within it. We would also know that the cultivation of vice is something that we should avoid. Even though it may be too early to talk about virtuous robots by design (Dignum 2018), the case study—involving social robots—we just reviewed shows that robots may be designed to promote and foster key human virtues.

Implementing the virtue ethics approach by both design and regulation means translating wisdom (phronesis) into rules that have exceptions, depending on the specific situation. For example, a vending machine will not dispense a bottle of water even if it is only a penny short of the required amount. For a penny, a person may die of thirst. A human operator can be more flexible. Of course, automation has many other advantages, one of which—as we have seen above—is greater efficiency than humans; however, it seems to be too strict to accommodate for the complexity of the real world.

To better illustrate this idea, we could perhaps discuss a more pertinent example. The example involves robot assistants for the elderly or for people with disabilities, which may be implemented to allow greater autonomy and agency for the individual. Let us imagine the case of an elderly diabetic with a non-severe form of the disease. She cannot eat sweets, as a general rule, but an occasional pastry will not be harmful. A robot assistant will have to prevent the elderly person from eating sweets, either

under the framework of deontological ethics (a prescription deduced from the behaviour that is right to follow) or under the framework of consequentialist ethics (the result will be better health). However, the comfort that an occasional pastry can provide to a depressed individual may be better illustrated through the lenses of a virtue ethics approach (we are not excluding that it may be included in the other approaches as well).

Naturally, the assistant robot could be programmed so as to grant × number of pastries in a given unit of time (a month, for example), in accordance with the health and mood conditions of the individual under its care. For further concessions, the robot should consult the physician or the supporting administrator. However, the rigid programming could also be subject to modification based on supervised or unsupervised learning. The robot could increase or narrow its basic range of pastries per month based on the real time evolution of the patient's health and mood conditions. This pattern of action, as noticed above, seems to be better described by the flexibility characterising a virtuous ethics approach. A virtuous ethics approach in promoting the idea that virtues are acquired by continuous habituation and perfected through practice emphasises their context-dependent nature. Hence, it introduces an element of flexibility, which seems to best capture human nature but also probably best reflects the nature of current AI technologies (based on reinforcement learning) that take flexibility as a core feature or trademark.

Embracing and defending a viewpoint grounded on virtue ethics (Bynum 2006), thus promoting the enhancement of human nature (Persson and Savulescu 2012) could respect human dignity and other crucial moral values, while contributing to reach *eudaimonia*; a state of human flourishing.

As Bynum (2006) and Stahl (2021) brilliantly pointed out, flourishing requires excellence skills in the pursuit of one's goals, which implies that—possibly—there are as many ways of flourishing as there are combinations of individual skills. This is an important point to emphasise for responsibilists like us (Karimov et al. 2022; Pietrini et al. 2022; Lavazza and Farina 2021), who think that there is a significant responsibility on the part of the individual in achieving virtuous behaviour. Nevertheless, it must also be noted that humans cannot flourish on their own. 'To live happy, meaningful lives, they must live together in communities—sharing experiences, challenges, common values—working to create or preserve a social context in which security, opportunities, knowledge, resources, and the other "core goods" are available in the community' (Bynum 2006, p.165). This therefore entails that the socio-cultural constructs and milieus surrounding the individuals (be them the state or the communities in which people are immersed) should be considered as equally important in defining, soliciting, and endorsing a virtuous-centered approach to AI.

5 Conclusion

As we saw from the discussion of the cases we described above (concerning workplaces, social robotics, administration of justice and health), the benefits related to the implementation of AI technologies in society may be unprecedented; however, the risks (linked to possible abuses of established standards and values) may also be significant.

It is quite possible that in the near future human beings will be deprived of their work and operational competence (at least in certain domains), that their companions will become intelligent robots, that they will no longer be able to enforce the attribution of criminal responsibility through their proxies (the judges), and that even medicine will become merely a depersonalized process. In such a scenario, individuals who do not need to work, who lose their main purpose of existence because almost everything is taken care of by automated systems, who see their direct interactions with their fellow human beings diminished, and who also find that the sphere of justice and medicine is administered by machines could—perhaps—achieve improvements in their material conditions; yet, they may not have a corresponding increase in their overall well-being.

We believe the major challenge of a good society in the age of AI lies precisely in reaping the greatest gains from intelligent automation, so as to remove as many people as possible from the realm of need and suffering, but also to prevent an epidemic of sadness, depression, and mental illness that may well arise as a consequence of such 'facilitated living'.

Data Availability Statement The manuscript has no associated data.

Declarations

Conflict of interest The authors assume all responsibility of the inception, development, writing, editing and final approval of the manuscript in its submitted form and have no conflicts of interest to declare.

References

Abeliansky A, Prettner K (2017) Automation and demographic change. Available at SSRN 2959977 518:1–44

Acemoglu D, Restrepo P (2020) Robots and jobs: evidence from us labor markets. J Polit Econ 128(6):2188–2244

Ahuja AS (2019) The impact of artificial intelligence in medicine on the future role of the physician. PeerJ 7:7702

Alemi M, Ghanbarzadeh A, Meghdari A, Moghadam LJ (2016) Clinical application of a humanoid robot in pediatric cancer interventions. Int J Soc Robot 8(5):743–759

Altman R (2015) Distribute AI benefits fairly. Nature 521(7553):417–418

Alvarado R (2021) Should we replace radiologists with deep learning? Pigeons, error and trust in medical AI. Bioethics 36:121–133
Anderson SL, Anderson M (2011) A prima facie duty approach to machine ethics and its application to elder care. In: Workshops at the twenty-fifth AAAI conference on artificial intelligence, pp 2–7
Arendt H (1950) The human condition. University of Chicago Press, Chicago
Arntz M, Gregory T, Zierahn U (2017) Revisiting the risk of automation. Econ Lett 159:157–160
Barrera E (2020) Technology and the virtues: a philosophical guide to a future worth wanting. Glob Med J 12(1):128–131
Bauer WA (2020) Virtuous vs. utilitarian artificial moral agents. AI Soc 35(1):263–271
Bench-Capon T, Araszkiewicz M, Ashley K, Atkinson K, Bex F, Borges F, Bourcier D, Bourgine P, Conrad JG, Francesconi E et al (2012) A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. Artif Intell Law 20(3):215–319
Blake W, Bloom H (1982) The complete poetry and prose of William Blake. University of California Press, Berkeley
Bostrom N (2017) Superintelligence. Oxford University Press, Oxford
Breazeal CL (2002) Designing sociable robots. MIT Press, Cambridge
Breazeal C, Dautenhahn K, Kanda T (2016) In: Siciliano B, Khatib O (eds) Social robotics, Cham. Springer, Berlin, pp 1935–1972. https://doi.org/10.1007/978-3-319-32552-1_72
Brooks RA (1999) Cambrian intelligence: the early history of the new AI. MIT Press, Cambridge
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, Mcafee A (2017) Artificial intelligence, for real. Harvard Business Review, Brighton
Buchanan BG (2005) A (very) brief history of artificial intelligence. AI Mag 26(4):53
Bugayenko Y, Bakare A, Cheverda A, Farina M, Kruglov A, Plaksin Y, Succi G, Pedrycz W (2022a) Automatically prioritizing and assigning tasks from code repositories in puzzle driven development. In: 2022 IEEE/ACM 19th international conference on mining software repositories (MSR). IEEE, pp 722–723
Bugayenko Y, Daniakin K, Farina M, Jolha F, Kruglov A, Succi G, Pedrycz W (2022) Extracting corrective actions from code repositories. In: 2022 IEEE/ACM 19th international conference on mining software repositories (MSR). IEEE, pp 687–688
Bush V et al (1945) As we may think. Atl Mon 176(1):101–108
Bynum TW (2006) Flourishing ethics. Ethics Inf Technol 8(4):157–173
Cath C, Wachter S, Mittelstadt B, Taddeo M, Floridi L (2018) Artificial intelligence and the 'good society': the US, EU, and UK approach. Sci Eng Ethics 24(2):505–528
Ciancarini P, Farina M, Masyagin S, Succi G, Yermolaieva S, Zagvozkina N (2021a) Non verbal communication in software engineering—an empirical study. IEEE Access 9:71942–71953
Ciancarini P, Farina M, Masyagin S, Succi G, Yermolaieva S, Zagvozkina N (2021b) Root causes of interaction issues in agile software development teams: status and perspectives. In: Future of information and communication conference. Springer, Berlin, pp 1017–1036
Clarke S, Zohny H, Savulescu J (2021) Rethinking moral status. Oxford University Press, Oxford
Coeckelbergh M (2020) AI ethics. MIT Press, Cambridge
Constantinescu M, Crisp R (2022) Can robotic AI systems be virtuous and why does this matter? Int J Soc Robot. https://doi.org/10.1007/s12369-022-00887-w
Constantinescu M, Voinea C, Uszkai R, Vică C (2021) Understanding responsibility in responsible AI. Dianoetic virtues and the hard problem of context. Ethics Inf Technol 23:803–814
Copeland BJ (2000) The turing test. Minds Mach 10(4):519–539
Cowls J, King T, Taddeo M, Floridi L (2019) Designing AI for social good: seven essential factors. https://doi.org/10.2139/ssrn.3388669
Danaher J (2019) The philosophical case for robot friendship. J Posthum Stud 3(1):5–24
Danziger S, Levav J, Avnaim-Pesso L (2011) Extraneous factors in judicial decisions. Proc Natl Acad Sci 108(17):6889–6892
Daugherty PR, Wilson HJ (2018) Human + machine: reimagining work in the age of AI. Harvard Business Press, Brighton
Davenport TH, Kirby J (2016) Only humans need apply: winners and losers in the age of smart machines. Harper Business, New York
de Fine Licht K, de Fine Licht J (2020) Artificial intelligence, transparency, and public decision-making. AI Soc 35(4):917–926
Dean J, Corrado GS, Monga R, Chen K, Devin M, Le QV, Mao MZ, Ranzato M, Senior A, Tucker P et al (2012) Large scale distributed deep networks. NIPs 2012(1):1223–1231
Delcker J (2018) Europe's silver bullet in global AI battle: ethics. https://www.politico.eu/article/europe-silver-bullet-global-ai-battle-ethics/. Accessed 4 Sept 2022
Dignum V (2018) Ethics in artificial intelligence: introduction to the special issue. Springer, Berlin
Douglas T, Pugh J, Singh I, Savulescu J, Fazel S (2017) Risk assessment tools in criminal justice and forensic psychiatry: the need for better data. Eur Psychiatry 42:134–137
Dreyfus H (1976) What computers can't do. Harper Collins, New York
Dreyfus HL, Hubert L et al (1992) What computers still can't do: a critique of artificial reason. MIT Press, Cambridge
Farina M, Lavazza A (2021a) Knowledge prior to belief: is extended better than enacted? Behav Brain Sci 44:e152. https://doi.org/10.1017/S0140525X2000076X
Farina M (2021b) Embodiment: dimensions, domains, and applications. Adapt Behav 29(1):73-99.https://doi.org/10.1177/10597123209129 6
Farina M, Lavazza A (2022a) Incorporation, transparency and cognitive extension: why the distinction between embedded and extended might be more important to ethics than to metaphysics. Philos Technol 35(1):1–21
Farina M, Lavazza A (2022b) Why there are still moral reasons to prefer extended over embedded: a (short) reply to Cassinadri. Philos Technol. https://doi.org/10.1007/s13347-022-00566-8
Felzmann H, Fosch-Villaronga E, Lutz C, Tamò-Larrieux A (2020) Towards transparency by design for artificial intelligence. Sci Eng Ethics 26(6):3333–3361
Floridi L (2019) Establishing the rules for building trustworthy AI. Nat Mach Intell 1(6):261–262
Floridi L (2020) What the near future of artificial intelligence could be. Philos Technol 32:1–15 https://doi.org/10.1007/s13347-019-00345-y
Floridi L, Cowls J (2022) A unified framework of five principles for AI in society. In Carta S (ed) Machine learning and the city: applications in architecture and Urban design. Wiley, Hoboken, NJ, pp 535–545
Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F et al (2018) AI4People-an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach 28(4):689–707
Fodor JA (1975) The language of thought, vol 5. Harvard University Press, Cambridge
Ford M (2017) Rise of the robots: technology and the threat of a jobless future. Basic Books, New York

Frey CB, Osborne MA (2017) The future of employment: how susceptible are jobs to computerisation? Technol Forecast Soc Change 114:254–280
Frisch M (1959) Homo faber. Abelard-Schuman, London
Gabriel I (2020) Artificial intelligence, values, and alignment. Minds Mach 30(3):411–437
Glöckner A (2016) The irrational hungry judge effect revisited: simulations reveal that the magnitude of the effect is overestimated. Judgm Decis Mak 11(6):601
Guo Z, Liu S, Liu J, Li Y, Chen L, Lu H, Zhou Y (2021) How far have we progressed in identifying self-admitted technical debts? A comprehensive empirical study. ACM Trans Softw Eng Methodol (TOSEM) 30(4):1–56
Hallamaa J, Kalliokoski T (2020) How AI systems challenge the conditions of moral agency? In: International conference on human–computer interaction. Springer, Berlin, pp 54–64
Heerink M, Vanderborght B, Broekens J, Albó-Canals J (2016) New friends: social robots in therapy and education. Int J Soc Robot 8(4):443–444
Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci 79(8):2554–2558
Isaac WS (2017) Hope, hype, and fear: the promise and potential pitfalls of artificial intelligence in criminal justice. Ohio State J Crim Law 15:543
Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399
Johal W (2020) Research trends in social robots for learning. Curr Robot Rep 1:75–83
Johnson DG, Verdicchio M (2019) AI, agency and responsibility: the VW fraud case and beyond. AI Soc 34(3):639–647
Jones SE (2013) Against technology: from the luddites to neo-luddism. Routledge, New York
Kahneman D (2011) Thinking, fast and slow. Macmillan Publishers, London
Karimov A, Lavazza A, Farina M (2022) Epistemic responsibility, rights, and duties during the Covid-19 pandemic. Soc Epistemol. https://doi.org/10.1080/02691728.2022.2077856
Kline R (2010) Cybernetics, automata studies, and the Dartmouth conference on artificial intelligence. IEEE Ann Hist Comput 33(4):5–16
Krakovsky M (2018) Artificial (emotional) intelligence. Commun ACM 61(4):18–19
Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
Kurzweil R (2005) The singularity is near: when humans transcend biology. Penguin Books, London
Lalsing V, Kishnah S, Pudaruth S (2012) People factors in agile software development and project management. Int J Softw Eng Appl 3(1):117
Lavazza A, Farina M (2021) Experts, naturalism, and democracy. J Theory Soc Behav 52(2):279–297
Le Q, Miralles-Pechuán L, Kulkarni S, Su J, Boydell O (2020) An overview of deep learning in industry. In: Liebowitz (ed) Data analytics and AI. CRC Press, Boca Raton, FL, pp 65–98
LeCun Y, Haffner P, Bottou L, Bengio Y (1999) Object recognition with gradient-based learning. In: Forsyth D, Mundy J, Gesu V, Cipolla R (eds) Shape, contour and grouping in computer vision, Lecture Notes in Computer Science. Springer, Berlin, Germany, pp 319–345
Li O (2021) Problems with "friendly AI''. Ethics Inf Technol 23(3):543–550
MacIntyre A (1981) After Virtue. University of Notre Dame Press, Notre Dame, IN
Markoff J (2016) Machines of loving grace: the quest for common ground between humans and robots. Ecco, New York
Marti P (2010) Robot companions: towards a new concept of friendship? Interact Stud Soc Behav Commun Biol Artif Syst 11(2):220–226
McCarthy J, Minsky M, Rochester N (1956) The Dartmouth summer research project on artificial intelligence. Artif Intell Past Present Future, pp 1–13. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf
McClelland JL, Rumelhart DE, Hinton GE (1986) The appeal of parallel distributed processing. MIT Press, Cambridge, pp 3–44
McCorduck P, Cfe C (2004) Machines who think: a personal inquiry into the history and prospects of artificial intelligence. CRC Press, Boca Raton
McKay C (2020) Predicting risk in criminal procedure: actuarial tools, algorithms, AI and judicial decision-making. Curr Issues Crim Justice 32(1):22–39
Megha S, Salem H, Ayan E, Mazzara M, Aslam H, Farina M, Bahrami MR, Ahmad M (2021) Survey on blockchain applications for healthcare: reflections and challenges. In: International conference on advanced information networking and applications. Springer, Berlin, pp 310–322
Minsky M, Papert S (1969) Perceptrons: an introduction to computational geometry. MIT Press, Cambridge
Mitchell RJ (1990) Managing complexity in software engineering. Peter Peregrinus, London
Montague PR, Dolan RJ, Friston KJ, Dayan P (2012) Computational psychiatry. Trends Cogn Sci 16(1):72–80
Montes GA, Goertzel B (2019) Distributed, decentralized, and democratized artificial intelligence. Technol Forecast Soc Change 141:354–358
Newell A, Simon HA et al (1972) Human problem solving. Prentice-Hall, Englewood Cliffs
Newell A, Shaw JC, Simon HA (1959) Report on a general problem solving program. In: IFIP congress, vol 256, p 64
Norvig PR, Intelligence SA (2002) A modern approach. Prentice Hall, Upper Saddle River
Nussbaum MC (1999) Virtue ethics: a misleading category? J Ethics 3(3):163–201
Palmer A, Schwan D (2021) Beneficent dehumanization: employing artificial intelligence and carebots to mitigate shame-induced barriers to medical care. Bioethics 36(2):187–193
Peeters A, Haselager P (2021) Designing virtuous sex robots. Int J Soc Robot 13(1):55–66
Persson I, Savulescu J (2012) Unfit for the future: the need for moral enhancement. Oxford University Press, Oxford
Philbeck T, Davis N (2018) The fourth industrial revolution. J Int Aff 72(1):17–22
Pietrini P, Lavazza A, Farina M (2022) Covid-19 and biomedical experts: when epistemic authority is (probably) not enough. J Bioeth Inq 19(1):135–142
Raisch S, Krakowski S (2021) Artificial intelligence and management: the automation–augmentation paradox. Acad Manag Rev 46(1):192–210
Renda A et al (2019) Artificial intelligence. ethics, governance and policy challenges. CEPS Centre for European Policy Studies, Brussels
Rosenblatt F (1960) Perceptron simulation experiments. Proc IRE 48(3):301–309
Rumelhart DE, Hinton GE, McClelland JL et al (1986) A general framework for parallel distributed processing. Parallel Distrib Process Explor Microstruct Cogn 1(45–76):26
Rumelhart DE, Widrow B, Lehr MA (1994) The basic ideas in neural networks. Commun ACM 37(3):87–93

Sand M, Durán JM, Jongsma KR (2021) Responsibility beyond design: physicians' requirements for ethical medical AI. Bioethics 36(2):162–169. https://doi.org/10.1111/bioe.12887
Savulescu J (2009) Moral status of enhances beings: what do we owe the gods? In: Savulescu J, Bostrom N (eds) Human Enhancement. Oxford University Press, Oxford, UK, pp 211–247
Savulescu J, Maslen H (2015) Moral enhancement and artificial intelligence: moral AI? In: Romportl J, Zackova E, Kelemen J (eds) Beyond artificial intelligence: the disappearing human-machine divide. Springer, Berlin, pp 79–95
Schuller D, Schuller BW (2018) The age of artificial emotional intelligence. Computer 51(9):38–46
Schwaber K (1997) Scrum development process. In: Sutherland D, Patel D, Casanave C, Hollowell G, Miller J (eds) Business object design and implementation. Springer, Berlin, pp 117–134
Searle JR (1982) The Chinese room revisited. Behav Brain Sci 5(2):345–348
Smuha NA (2019) The EU approach to ethics guidelines for trustworthy artificial intelligence. Comput Law Rev Int 20(4):97–106
Spiekermann S, Krasnova H, Hinz O, Baumann A, Benlian A, Gimpel H, Heimbach I, Köster A, Maedche A, Niehaves B et al (2022) Values and ethics in information systems. Bus Inf Syst Eng 64(2):247–264
Spinellis D (2005) Version control systems. IEEE Softw 22(5):108–109
Stahl BC (2021) Concepts of ethics and their application to AI. In: Stahl BC (ed) Artificial Intelligence for a Better Future, Springer Briefs in Research and Innovation Governance. Springer, Cham, pp 19–33
Sternberg RJ (1983) Components of human intelligence. Cognition 15(1–3):1–48
Storey M-A, Ryall J, Bull RI, Myers D, Singer J (2008) Todo or to bug. In: 2008 ACM/IEEE 30th international conference on software engineering. IEEE, pp 251–260. https://doi.org/10.1145/1368088.1368123
Susskind RE, Susskind D (2015) The future of the professions: how technology will transform the work of human experts. Oxford University Press, Oxford
Taddeo M, Floridi L (2018) How AI can be a force for good. Science 361(6404):751–752
Trappl, R (2015) A construction manual for robots' ethical systems. Springer, Berlin
Turing AM, Haugeland J (1950) Computing machinery and intelligence. MIT Press, Cambridge
Vallor S (2016) Technology and the virtues: a philosophical guide to a future worth wanting. Oxford University Press, Oxford
Vallor S (2017) AI and the automation of wisdom. In: Powers T (ed) Philosophy and Computing Essays in Epistemology, Philosophy of Mind, Logic, and Ethics. Springer, Berlin, pp 161–178
Wajcman J (2017) Automation: is it really different this time? Br J Sociol 68(1):119–127
Wallach W, Vallor S (2020) Moral machine: from value alignment to embodied virtue. In: Liao M (Ed). Ethics of Artificial Intelligence. Oxford University Press, New York, NYC, pp 383–412
Walsh T, Levy N, Bell G, Elliott A, Maclaurin J, Mareels I, Wood F (2019) The effective and ethical development of artificial intelligence: an opportunity to improve our wellbeing. Australian Council of Learned Academies, Melbourne
Wang R (2020) Legal technology in contemporary USA and China. Comput Law Secur Rev 39:105459
Wasserman AI (1996) Toward a discipline of software engineering. IEEE Softw 13(6):23–31
Weidong J (2020) The change of judicial power in China in the era of artificial intelligence. Asian J Law Soc 7(3):515–530
Weizenbaum J (1976) Computer power and human reason: from judgment to calculation. WH Freeman & Co, New York
Wiener N (1988) The human use of human beings: cybernetics and society. Da Capo Press, Boston
Wiese W, Friston KJ (2021) Ai ethics in computational psychiatry: from the neuroscience of consciousness to the ethics of consciousness. Behav Brain Res 420:113704
Wright SA, Schultz AE (2018) The rising tide of artificial intelligence and business automation: developing an ethical framework. Bus Horiz 61(6):823–832
Xia Q, Sifah EB, Asamoah KO, Gao J, Du X, Guizani M (2017) Medshare: trust-less medical data sharing among cloud service providers via blockchain. IEEE Access 5:14757–14767
Yang G, Bellingham J, Dupont P, Fischer P, Floridi L, Full R, Jacobstein N et al (2018) The grand challenges of science robotics. Sci Robot 3(14):eaar7650