AI and Ethics
https://doi.org/10.1007/s43681-022-00222-z

ORIGINAL RESEARCH

"AI for all" is a matter of social justice

Alessandra Buccella[1]

Received: 6 May 2022 / Accepted: 9 September 2022

Abstract
Artificial intelligence (AI) is a radically transformative technology (or system of technologies) that created new existential possibilities and new standards of well-being in human societies. In this article, I argue that to properly understand the increasingly important role AI plays in our society, we must consider its impacts on social justice. For this reason, I propose to conceptualize AI's transformative role and its socio-political implications through the lens of the theory of social justice known as the Capability Approach. According to the approach, a just society must put its members in a position to acquire and exercise a series of basic capabilities and provide them with the necessary means for these capabilities to be actively realized. Because AI is re-shaping the very definition of some of these basic capabilities, I conclude that AI itself should be considered among the conditions of possession and realization of the capabilities it transforms. In other words, access to AI—in the many forms this access can take—is necessary for social justice.

Keywords Artificial intelligence · Technology · Society · Capabilities · Justice · Nussbaum

> *…to secure a right to citizens in [a certain area] is to put them in a position of capability to function in that area. To the extent that rights are used in defining social justice, we should not grant that the society is just unless the capabilities have been effectively achieved.—Martha Nussbaum (24, 37).*

1 Introduction

In this article, I argue that, because AI is a radically transformative technology (or system of technologies) that created new possibilities and new standards of well-being in society, access to AI should be provided to all members of society. In this sense, the advent of AI and its applications should be understood in analogy with large-scale disruptive events such as the agricultural or the industrial revolution, which themselves re-shaped human needs, goals, freedoms, and opportunities. However, conceptualizing and properly understanding the sense in which social justice requires access to AI—at least in those societies where AI-powered technologies are routinely deployed and used—is no easy task. It requires not just a precise understanding of the very notion of AI, but also a theory of social justice that can accommodate the kind of radical transformations operated by AI on people's lives. I explore these issues through the lens of the theory known as the Capability Approach (e.g., [24, 25, 33–35].[1] According to this theory, the necessary (and possibly sufficient) conditions for social justice consist in a "political, social, and economic environment" [25], 20) in which a person is granted a set of *capabilities*. In turn, to be granted a capability means to have the opportunity to select and choose one's own functionings—i.e., the "beings and doings that are the outgrowths or the realizations of capabilities." (25). Importantly, being able to choose one's own functionings within the capability framework entails being able to choose functionings that explicitly reject or contradict the capabilities. To be granted a capability in a society, then, one must be provided with the necessary means for this capability to be either actively realized or deliberately rejected. In the remainder of the paper, I will refer to the possibility to reject or refuse to realize a capability as the "opting-out" aspect of capabilities, which is

✉ Alessandra Buccella
buccella@chapman.edu

[1] Institute for Interdisciplinary Brain and Behavioral Sciences, Chapman University, Orange, CA, USA

[1] See also Oosterlaken & van den Hoven [26].

Published online: 22 September 2022

complementary to, and just as important as, the positive realization aspect. Only by being put in a position to either realize or "opt out" of a capability, the proponents of the approach argue, people can live a life worthy of human dignity, which the approach sees as the basic condition for social justice [25]. For the purposes of this paper, I will assume that human capabilities, as understood by the capability approach, can be re-defined and their conditions of possession and realization can be modified by new technologies, as long as those technologies are disruptive and impactful enough. Even though I will not explicitly argue for this claim, I will nonetheless use the first part of the paper to defend the related idea that AI is the kind of technology—or rather, family of technologies—that can in fact transform human capabilities similar to how other radically transformative technologies have done in the past.

Thus, the first part of my argument is as follows:

P1. Radically transformative technologies change the definition and the conditions of possession and realization of at least some capabilities.
P2. With its mere existence, AI changes the definition and the conditions of possession and realization of at least some capabilities.
C1. Therefore, AI is a radically transformative technology.

Sections 2 and 3 will be dedicated to defending P1 and P2, respectively. Once C1 is established, I will proceed to arguing for the main thesis of the paper: namely, that access to AI, in a sense that will become clear as the second half of the argument unfolds, is required for social justice (defined according to the capability approach). The second part of the argument looks like this:

P3. A society is just if the fundamental human capabilities are granted to every member of that society.
P4. For a capability to be granted, its conditions of possession and realization (including opting-out conditions) must be met.
P5. For the conditions of possession and realization of a capability (including its opting-out conditions) to be met, practical and intellectual access to such conditions must be provided.
P6. One way in which a radically transformative technology changes the definition and the conditions of possession and realization of a capability is by itself becoming one of such conditions.
P7. For the capabilities radically transformed by AI to be granted, practical and intellectual access to AI must be provided.
C2. Therefore, a society whose capabilities (at least some of them) have been radically transformed by AI is just only if practical and intellectual access to AI is provided to all its members.

I take P3 and P4 to follow directly from the capability approach, which is the framework I am assuming. I will, however, say more to support P5 and P6 (§4), mostly with respect to the notion of "practical and intellectual access" and what it would look like when applied to AI as I have defined it.

2 Part I: the radically transformative nature of AI

Gruetzemacher and Whittlestone [17] distinguish three main classes of transformative effects that technology has on society:

1. A technology becomes so widespread and multi-purpose that it penetrates even the smallest aspects of life. These are called "General Purpose Technologies" (GPTs). Examples: telephone, electricity.
2. A technology has a quick and dramatic impact on a small but important aspect of life. Example: Nuclear energy (and the consequent availability of nuclear weapons) impacting warfare and international relations.
3. A technology indirectly and over a longer period of time precipitates "fundamental and unprecedented societal change" through "temporal clusters of technological innovation". Examples: the first industrial revolution, the agricultural revolution.

According to Gruetzemacher and Whittlestone's own terminology, all three types of transformative technologies lead to "practically irreversible change in trajectories of human life and progress" (2022, p. 7). In addition, once a trajectory has been established through practically irreversible changes, a further transformative effect allows the trajectory to be "locked in" and to endure over a long period of time. However, something even more radical than practically irreversible change and lock-in of trajectories seems to be involved in type-3 transformations. To see this clearly, though, it is first necessary to clarify further in what sense the agricultural and industrial revolutions can be themselves seen as technologies.

In his (1985b), Amartya Sen takes inspiration from Marx's definition of technology as "the combining together of various processes into a social whole", and emphasizes how the definition of technology should not be limited to "particular mechanical or chemical processes used in making one good or another." Rather, technology has a fundamentally *social* content, too, which should be acknowledged in the discussion around how to best respond to

technological transformations and how to make sure that technology is overall beneficial to humanity. According to Sen, "the making of things involves not merely the relationship between, say, raw materials and final products, but also the feasibilities of harnessing, utilizing and transforming raw materials into commodities through socially viable labour use and organization." The transformative power of technology, in other words, does not only involve "equipment and its operational characteristics, it is also about the social arrangements that permit productive processes to be carried out" [34], 2–3).

Because, as Sen argues, technology in general includes an irreducible social component, we can see technology which is radically transformative in Gruetzemacher and Whittlestone's sense as effectively creating a *new* society in which humans acquire new roles, conceive of their identity differently, face new responsibilities and opportunities, etc. In this sense, therefore, the agricultural and industrial revolutions qualify as technologies of the radically transformative kind:

> Both revolutions constituted extreme and unprecedented changes to human life: a transition from people living as hunter-gatherers to large, settled civilizations; and a transition to mechanized manufacturing and factories, leading to unprecedented population growth and rising quality of life. The industrial revolution in particular coincided with clear trajectory changes in metrics of human well-being including measures of physical health, economic well-being, energy capture and technological empowerment. [17], 5)

Just like these revolutions consisted in "temporal clusters of technological innovation" which brought about radical societal change over time, so I claim AI is doing right now (and has been doing for the past few decades). Machine learning, large language models, computer vision, and robotics are good examples of such temporal clusters within the family of technologies I refer to as AI.[2] As Sen pointed out, technology has a "social content" not only in the sense that it changes the material conditions of human life (by, for instance, introducing new equipment for the production of goods), but also, and more importantly, in the sense that it re-structures the social conditions in which equipment is used and goods are produced. Again following Sen, by social conditions I mean the social arrangements necessary to successfully operate the new equipment and handle its products. These social arrangements, in turn, include things like the specification of life quality standards for people involved in production and distribution of the goods, laws and regulations for labor, commerce, property ownership, etc., as well as the formulation of basic social, civil, and political rights. Even before one looks at the capability approach as a full-blown, predictive theory of social justice, one will notice how the approach is first and foremost an attempt to systematize these social arrangements and, consequently, to give a society that is going through a time of technological change the conceptual tools to re-structure itself. Consistently with this framework, therefore, I suggest that a radically transformative technology in Gruetzemacher and Whittlestone's sense is one that changes the "social arrangements" of the society it is introduced into, that is, it changes the conditions of possession and realization of capabilities within that society. In the next section, I argue that AI qualifies as a radically transformative technology in this sense, since through different technological innovation "clusters", such as the development and deployment of machine learning, robots, language models, etc. it is changing the conditions of possession and realization of at least some capabilities.

3 The radically transformative effects of AI on capabilities

Initially, the capability approach was proposed as a way to assess a society's quality from a standpoint different from the standard, average income-based one [35]. However, according to Nussbaum, the approach had the potential to do much more than simply suggesting a new conceptual framework: it could provide a series of concrete criteria to evaluate how just a society is, and to identify possible areas of improvement to tackle through legislation and political advocacy. Nussbaum's view of capabilities as "fundamental entitlements", thus, leads to the formulation of what she considers the ten "central capabilities" [24], 41–42).[3] Here I mainly discuss seven of Nussbaum's ten capabilities, namely, those that I consider most radically redefined by AI and which, therefore, now have AI as one of their conditions of possession and realization. I do not have a principled criterion to determine at what point and in virtue of what exactly a capability has been thoroughly transformed by AI. However,

[2] On page 12 of their article, Gruetzemacher and Whittlestone write: "there is no clear evidence or consensus that any single technology has alone precipitated change on the level we are describing as "radical societal transformation"—historically these changes seem to have resulted from clusters of technologies potentially in interaction with other societal factors. However, AI is arguably unique in that it does not necessarily represent a single technology, but an underlying method leading to a cluster of different technologies: including, for example, natural language processing, computer vision, and robotic learning. Thus, […] it is also possible that a cluster of different AI technologies could lead to TAI [Transformative AI] or RTAI [Radically Transformative AI]".

[3] I will not discuss all ten capabilities. In particular, I will not explicitly mention in the paper the capabilities called Emotions, Practical Reason, and Play. See Nussbaum [24] for more on those.

for each capability I give some suggestive examples which, taken collectively, I believe can make up for the absence of a precise criterion, at least for the purposes of this paper.

> **Life.** Being able to live to the end of a human life of normal length; not dying prematurely, or before one's life is so reduced as to be not worth living.

The first capability has been affected by the advent of AI in our society in several ways already. For example, machine learning algorithms are now routinely being employed in natural disaster prediction, prevention, and to coordinate evacuation and rescue operations. In war zones, AI can be employed to guide missile strikes toward strategic objectives and to minimize civilian casualties. Thus, AI-powered technologies allow for more human (and non-human) lives to be saved, for these lives to last longer, and to not be prematurely interrupted.

> **Bodily Health.** Being able to have good health, including reproductive health; to be adequately nourished; to have adequate shelter.

Neural networks trained on big data can now be used in a variety of healthcare-related domains, for instance to issue recommendations regarding resource management in hospitals (bed allocation, ventilators and PPE availability, etc.), transplant waitlists, and more [3]. Some surgeries and exploratory exams can now be performed by high-precision robots aided by machine learning ([16, 19], and AI systems that issue medical diagnoses and suggest treatments have been around for a while [11, 18]. Some have explicitly argued that machine learning and computational AI more generally are transforming psychiatry by changing the conceptual categories the field relies on, and might soon transform the very definition of mental disorder [42]. Finally, machine learning is also used in architecture and building engineering to create safer, more affordable, and more sustainable housing (e.g., [20, 22, 23].

> **Bodily Integrity.** Being able to move freely from place to place; to be secure against violent assault, including sexual assault and domestic violence; having opportunities for sexual satisfaction and for choice in matters of reproduction.

AI is also at the center of major advancements in the transportation industry, including self-driving vehicles, design of more efficient interchanges and highway routes, and elaboration of enormous amounts of data regarding people's movement patterns, habits, preferences, etc. collected through exercise apps, travel websites, public transportation online platforms, navigation software like Google Maps, etc. (see [1] for a review). With respect to reproduction, AI is instrumental in the collection and elaboration of related data through menstrual cycle and ovulation tracking, digital pregnancy tests, prescription-free online purchase of contraceptives, etc. AI is also being employed more and more in the context of fertility medicine (e.g., [9, 12, 37]. Protection from violence, sexual or of other nature, is also something that is being re-defined by AI technologies, such as 'intelligent' house alarm systems that communicate directly with law enforcement or automatically lock doors and windows when a home invasion attempt is detected.

Finally, machine learning algorithms that process data collected through dating apps and social media are now able to provide individuals with suggestions of events, activities, and even specific partners tailored to their sexual preferences and identities.

> **Senses, Imagination, and Thought.** Being able to use the senses, to imagine, think, and reason—and to do these things in a "truly human" way, a way informed and cultivated by an adequate education, including, but by no means limited to, literacy and basic mathematical and scientific training. Being able to use imagination and thought in connection with experiencing and producing works and events of one's own choice, religious, literary, musical, and so forth. Being able to use one's mind in ways protected by guarantees of freedom of expression with respect to both political and artistic speech, and freedom of religious exercise. Being able to have pleasurable experiences and to avoid nonbeneficial pain.

The major role AI plays in the context of education, entertainment, and other related fields is quite easy to identify. AI and machine learning applications are transforming work and education in a number of ways: from remote schooling databases to hybrid work platforms, from entirely online higher education programs to language-learning apps. Consider also sport and exercise technology like the Tonal home gym (with AI-powered adaptive workouts and digital weight), whole-body gaming devices (equipped with sophisticated computer vision software) like Kinect, or the cultural and recreational experiences available in virtual reality.

A further interesting case is represented by AI-relying accessibility features on personal devices, such as speech-to-text, text-to-speech, color-blind display mode, handwriting recognition, variable haptic feedback, enlarged fonts, etc. I include these features under this capability not with the intent of suggesting that what now counts as "doing things in a truly human way" is defined by AI and can only be accomplished through AI or with the help of AI. The availability of AI-powered tools does not by itself have the capacity to make us more or less human. Every human is "truly" human, whether or not it uses AI-powered tools. What I do want to claim, instead, is that the fact that AI-powered accessibility features exist puts people in a position to choose for themselves what "doing things in a truly human way" means

to them, being less limited by certain practical constraints connected with their physical or mental abilities, preferences, and inclinations. A world in which people are given the opportunity to explore and choose among many different versions of "doing things in a truly human way" is a world in which the fourth capability is more likely to be realized.

Affiliation

A. Being able to live with and toward others, to recognize and show concern for other human beings, to engage in various forms of social interaction; to be able to imagine the situation of another. (Protecting this capability means protecting institutions that constitute and nourish such forms of affiliation, and also protecting the freedom of assembly and political speech.)

B. Having the social bases of self-respect and nonhumiliation; being able to be treated as a dignified being whose worth is equal to that of others. This entails provisions of nondiscrimination on the basis of race, sex, sexual orientation, ethnicity, caste, religion, national origin.

In this domain, one immediately thinks about social media (one of the first "mass" applications of AI technology) and the role they have been playing for the past two decades or so. The very notion of affiliation, the processes through which people form and cultivate relationships, and the conceptual frameworks that allow people to understand such relationships, have been shaped by AI. Preferences, expectations, conceptions of what we need and what we deserve from a friend, a partner, a colleague, an employer, an employee—in fewer words, our social goals and social norms—have all been re-defined by AI applications. Even more so since the coronavirus pandemic started, access to products like social networks, dating apps, or videocall platforms is among the conditions of possession and realization of the Affiliation capability. Machine learning algorithms able to extract patterns from social media use and issue recommendations of websites, events, etc. are also changing the way we construct and connect the various aspects of our identities, how we engage in politics, how we develop empathy, and how we maintain emotional connections to others. AI-powered tools make it easy to share, and the more aspects of life are shared, the more one will find others responding to and engaging with them. Of course, with AI redefining the meaning of notions like "socialization", "relationship", or "shared experience" come an entirely redefined set of threats to our social well-being. For instance, many people who regularly use dating apps experience emotional burnout-like symptoms due to the potentially infinite number of "matches", the repetitiveness of how these connections evolve, especially in the early stages, and the pressure to keep up with all of them in a constant state of anticipation.[4]

Other Species. Being able to live with concern for and in relation to animals, plants, and the world of nature.

I have pointed out earlier that a capability can be considered "transformed by AI" when AI changes the conditions of its possession and realization. While AI might not have completely transformed this capability yet, I think that we have enough evidence suggesting that it might do so in the not-so-distant future. In particular, AI is starting to show a lot of potential in areas like wildlife conservation and the preservation of biodiversity [41]. For example, databases of drone and satellite images can help track and categorize endangered animal species down to the individual animal, helping re-population efforts and the fight against poaching. In addition, machine learning algorithms can offer insights and solutions into the trends of certain animal populations, their movements, habits, and preferences, as well as mapping food availability and predict migration routes. AI-equipped technologies are also employed in the fight against climate change [31] and in the development of alternatives to fossil fuels [21]. These examples of AI applications suggest that having the capability to "live with concern for and in relation to nature" one day might entail the ability to make decisions based on data collected through these AI-powered methods, thus changing the standards for a fulfilling relationship with nature and the expectations about what we can and ought to do to make it even more fulfilling.

However, it is worth noting that this coin has another, "darker" side: AI can have high negative impact on the world of nature, too. For example, high-tech, AI-powered telescopes and supercomputers used in astronomy have a quite large carbon footprint, and the energy costs of training large artificial neural networks are very high [2, 28, 36]. Thus, more has to be said about how we can protect ourselves against the negative effects of the radical changes operated by AI on capabilities.

Control Over One's Environment

A. Political. Being able to participate effectively in political choices that govern one's life; having the right of political participation, protections of free speech and association.

B. Material. Being able to hold property (both land and movable goods), and having property rights on an equal basis with others; having the right to seek

[4] For more on this, see https://www.nytimes.com/2022/08/31/well/mind/burnout-online-dating-apps.html?smid=url-share.

> employment on an equal basis with others; having the freedom from unwarranted search and seizure. In work, being able to work as a human being, exercising practical reason, and entering into meaningful relationships of mutual recognition with other workers.

I will use this last capability to say more about how AI is transforming or has already transformed the definition and the conditions of possession and realization of (some) capabilities in ways that are not all positive. I want to draw attention to the fact that AI technology can, in addition to expanding and making capabilities easier to possess and realize, also make them harder to possess and realize, due to the dangers that inevitably accompany a family of technologies that we do not (yet) fully understand or control. As much as AI has re-defined the conditions of possession and realization of at least some central capabilities, it has also re-defined the ways in which those capabilities can be taken away. To give one example, the advent of big data analysis has made seeking and obtaining employment a more standardized process (most first-round selections are done by algorithms based on keywords and other information contained in digital resumes). Although this process has the potential to increase objectivity and to pair employers and employees more effectively, it might make it easier to "get away with" (voluntary or involuntary) discrimination. Because of how advanced natural language processing is today, algorithms have the power to retrieve information about a potential hire's personal life, race, cultural and religious background, sexual orientation, and more that might be implicitly present in one's application documents.

As a second example, consider how AI-powered technologies have profoundly altered what it is to "hold property" and possibly the very concept of property itself. Stock markets are fully digitalized and are now "located" inside super-computers. Powerful machine learning algorithms mine data from all over the planet to predict trends and significant events that can affect local and global economies. The very existence of this complex, largely opaque, and evasive system comes with an exponentially increased risk of sudden and unpredictable financial, political, and social collapses around the world.

How can these two aspects be reconciled? First, notice that an essential feature of the capability approach articulated by Nussbaum is that one must be free to *choose whether or not to realize the capabilities one is granted*. This requirement is important, because it avoids the risk of just societies becoming paternalistic societies, in which the government positively dictates to its citizens how to act and what to choose to live "the right way". One essential feature of capabilities is, on the contrary, their openness to being deliberately ignored or rejected by individuals: "to promote capabilities", Nussbaum writes, "is to promote areas of freedom, and this is not the same as making people function in a certain way" (2011, 25). This aspect becomes even more central for AI-transformed capabilities, since some of those capabilities enable functionings that are controversial at best, and sometimes unambiguously bad for certain members of society or social groups. A society that promotes capabilities must promote ways for people to *both* function in accordance with these capabilities and opt out of those functionings if they so want. In other words, it is part of the very definition of a capability that one is allowed not to take up the corresponding functioning, and that this right to "opting out" is enforced and protected. Therefore, insofar as AI transforms the conditions of possession and realization of a capability, AI must guarantee that the conditions for opting out are met, too. For instance, AI-powered cybersecurity and countersurveillance technology can protect people's privacy and compensate for the risks that come from the massive and largely uncontrolled flow of information that fuels machine learning technology. However, if protection from AI-generated threats to AI-transformed capabilities comes from AI use itself, then it seems that AI is necessary not only to possess and realize capabilities, but also to opt out of them. The second part of the paper is dedicated to defending this claim via the key notion of "practical and intellectual access" to AI.

4 Part II: the role of practical and intellectual access to AI

At this point of the discussion, one might point out that AI-powered technologies are impacting society in a way that is not radically transformative and definitely less systemic that my view makes it sound. For quite some time now, supporters of the so-called Extended mind theory [5–8] have been arguing that new AI-equipped tools like robotic prosthetic devices, smartphones, fitness watches, etc. have the power to integrate and modify cognitive processes like memory, sensory perception, and even logical reasoning. AI and machine learning are also integrated into deep-brain stimulation implants used to alleviate symptoms in neurodegenerative diseases like Parkinson's [4, 10], and in sensory substitution and restoration devices to produce more accurate and fulfilling sensory experiences (e.g., [29, 40]. Without appealing to the complex framework of the capability approach, the extended cognition hypothesis seems nonetheless able to account for the role AI plays in changing the definition of a functioning human being.

However, as soon as one tries to unpack the details of the extended cognition hypothesis and makes its implications explicit, multiple problems emerge. Some of these problems, like that concerning how to exactly interpret the claim that non-biological tools or processes are constitutively part of

cognition, have been known since the early days of the proposal [32]; others have gained traction more recently as they apply specifically to state-of-the-art AI technology, including deep neural networks [38]; others again derive from the difficult task of constructing a new set of epistemic norms that work for artificially augmented or 'extended' knowers (e.g., [27, 30]. More generally, the extended cognition hypothesis seems committed to a transformative role for AI comparable to that of general-purpose technologies, that is, a 'level-1' transformation according to Gruetzemacher and Whittlestone's proposal (cfr. §2). In turn, this means that in the extended cognition picture AI falls short of changing the definition and the conditions of possession and realization of capabilities. In the language of the capability approach, all AI does according to the extended cognition hypothesis is provide new material routes to the possession and realization of capabilities which, however, remain the same at their core. The key difference between my proposal and the extended cognition hypothesis, therefore, is the following: on my view, AI does more than helping particular people acquire and exercise certain capabilities. It changes the way in which we determine whether a capability is acquired and exercised *by anyone* in a given society.

The shortcomings of the extended cognition perspective on AI's role in society allow me to indirectly connect back to the next step in my argument: once a technology becomes one of the conditions of possession and realization of a capability (as well as what guarantees the possibility of opting out, as I argued earlier), a society is required to provide access to that technology to all members if it wants to claim that the capability in question is granted in that society. After giving a tentative definition of access in terms of "practical and intellectual" access, I will proceed to explain how practical and intellectual access to AI is a requirement for the 'positive' realization of AI-transformed capabilities on the one hand, and for their 'negative' opting-out aspect, on the other.

Consider the following analogy. An alien with a surprisingly human-like body comes to Earth and wants to live here. In the city the alien landed in, the only existing means of transportation are bicycles. We might thus say that, to function in that community, one must have the capability of riding a bike. How does one acquire and realize the capability of riding a bike? It seems to me plausible to say that, minimally, the conditions of possession and realization of that capability include having—at least potential—access to a bike *and* to some theoretical knowledge about what bikes are and how they work. By having "practical" access to the conditions of possession and realization of a capability, therefore, I mean being in a position to use the material tools necessary to implement such conditions—if any. In the analogy above, practical access consists in the possibility of owning, renting, borrowing, etc. an actual bicycle. On the other hand, having "intellectual" access to the conditions of possession and realization of a capability entails being in a position to acquire relevant theoretical knowledge about the capability and its conditions of possession and realization. Continuing with the bicycle analogy, intellectual access will consist in the possibility of learning facts about bikes. I specifically say that one must be *in a position to acquire* practical and intellectual access to the relevant tools and facts, not that one must, in fact, access such tools and facts. This phrasing is crucial, as it preserves individuals' freedom to *not* realize a capability (recall Nussbaum's conception of capabilities as "areas of freedom" and the notion of "opting out"). While a society (and in particular its governing authorities) cannot force anyone to realize a capability, they must grant everyone with the possibility to do so, as well as the possibility to actively opt out. It is not, therefore, an individual responsibility of each member of society to put themselves in a position to gain practical and intellectual access to the conditions of possession and realization of a capability or to explicitly reject such conditions. It is the job of a society's governance to promote and maintain its members' practical and intellectual positioning with respect to the capabilities it grants. To go back to the bike analogy one last time, a society in which riding bikes is granted as a fundamental capability is a society in which the designed governing body provides bikes and "bike-related education" for free to everyone who might want them, while at the same time making sure that those who do not want them are not forced to have them to live life with dignity (for instance, by making sure that nobody is punished or discriminated against due to not wanting to ride bikes).[5]

Through the bike analogy, I aimed, on the one hand, to clarify the notions of practical and intellectual access to a technology (or family of technologies) and, on the other

[5] I am aware that the bike analogy is quite shallow compared to other more relevant and more significant areas of freedom societies are grappling with nowadays, such as wearing face masks to prevent the spread of COVID-19. While in this work I am not explicitly defending a specific position on issues of that sort, I want to make clear to the reader that I do *not* endorse a view that puts individual freedom above every other societal value without qualification. To claim that people should not be discriminated against or punished because they choose not to ride bikes is of course different from claiming that people should not be discriminated against or punished because they choose not to wear a face mask during a pandemic. Although I am not an expert in public policy, I am tempted to say that the *reasons* why a society promotes certain behaviors and the *goals* that a society is pursuing through the promotion of such behaviors matter when ensuring access to the conditions of possession and realization of a capability or the possibility to opt out. The *definition* of the capability in question matters, too, to determine what exactly its conditions of possession and realization are, and what "opting out" might look like. In sum, the reader should not take my examples and analogies to fully reflect my broader socio-political positions regarding individual freedom.

hand, to emphasize the importance of governing authorities whose job is to ensure both the functioning and the opting out aspects of capabilities by being in charge of their conditions of possession and realization. When AI becomes the technology in question, however, things need to be unpacked yet a little further. To do so, I now introduce another theoretical framework, this time specifically tailored to the role of AI in society, which proposes a series of guiding principles for developing, deploying, and using AI technologies in ethical and responsible ways [13, 14, 39]. According to this latter framework, there are five principles that inspire the creation of ethical and trustworthy AI.[6] AI technologies should be beneficent, non-maleficent, respect human autonomy, ensure and possibly promote justice, and be explicable. In particular, the principle of Autonomy states that "the autonomy of humans should be promoted and that the autonomy of machines should be restricted and made intrinsically reversible, should human autonomy need to be protected or re-established" [14], 7). It is useful, I think, to look at this principle to better understand the relationship between the requirement for practical and intellectual access to AI and the opting-out aspect of AI-transformed capabilities.

Consider another example. Machine learning algorithms can be used in healthcare to recommend diagnoses and treatments. However, these machines can only issue recommendations based on what they 'learn' from the databases used to train them, and, unfortunately, often these databases are affected by the same biases that have historically affected medical research and literature (since the data come precisely from there). Using big data for diagnostic purposes is, therefore, quite problematic, especially when members of social minorities are involved. In a recent paper, 15, 2) argue that "the very definition of a condition or disease hinges on gender or ethnicity, and has been used in a discriminatory fashion. Even for conditions that do not rely on race or gender, the gender- or race-specific presentation of a condition may be poorly understood, or ignored, in medical education and literature."

As I have argued previously (§3), the very idea of being in good health, i.e., one of Nussbaum's ten fundamental capabilities, has been transformed by big data and the AI-powered techniques used to analyze such data. With more information, faster ways to process it, and more efficient methods for finding significant patterns and trends, the standards for being in good health have significantly shifted. However, consider the case of someone belonging to an ethnic minority who is recommended by their AI-informed doctor to follow a therapy that was proven very effective according to the statistics calculated by the machine. Having the possibility to access the demographic information contained in the database the machine used to issue its recommendation, as well as to gain some general knowledge of algorithmic bias and its risks, seems in this case necessary in order for the good health capability to be granted to that patient. Indeed, the AI-transformed capability for good health requires that a patient is put in a position to either choose to trust and follow the machine's recommendations or refrain from consulting the machine altogether. As for the bike case, these two alternative options are granted only if the patient is put in a position to gain practical and intellectual access to the machine and its processes. The possibility to gain knowledge (both theoretical and practical) must be there regardless, to guarantee that the patient makes an informed, free, and autonomous decision even when opting out.

Clearly, the most direct way to ensure the opt-out possibility on the basis of practical and intellectual access to AI would be for the patient to gain knowledge of things like algorithmic bias and access to demographic information in the database. However, as I have argued above, a society is only required to put individuals *in a position to* have access to the conditions of possession and realization of a capability, while at the same time being in charge of such conditions through the relevant governing authority. In order not to force individual patients to themselves gain practical and intellectual access to the relevant AI technologies, there must be a system of governance in possession of the relevant knowledge and with access to the relevant material tools (which entails that this system must also be powered by AI technology) with the task of making sure that opting out is in fact a live possibility for the patient, independently of what the patient chooses. It is through the role of governing authorities, and not just individuals, that practical and intellectual access to AI is established as necessary for AI-transformed capabilities to be fully granted (in both their functioning and opting out aspects).

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

[6] The Capability Approach, in the words of its own founders, "is influenced by philosophical views that focus on human flourishing or self-realization, from Aristotle to John Stuart Mill in the West and Rabindranath Tagore in India." [25], 23). Similarly, the proponents of the "AI4people" framework emphasized that truly ethical and trustworthy AI should offer opportunities "for fostering human dignity and promoting human flourishing" [13], 690). Because of this shared insight, I believe that the two frameworks can be fruitfully combined even beyond the scope of the present work in order to shed more light on the socio-political implications of AI.

References

1. Abduljabbar, R., Dia, H., Liyanage, S., Bagloee, S.A.: Applications of artificial intelligence in transport: an overview. Sustainability (2019). https://doi.org/10.3390/su11010189
2. Aujoux, C., Kotera, K., Blanchard, O.: Estimating the carbon footprint of the GRAND project, a multi-decade astrophysics experiment. Astroparticle Phys. **131**, 102587 (2021). https://doi.org/10.1016/j.astropartphys.2021.102587
3. Awad, E., Levine, S., Anderson, M., Anderson, S.L., Conitzer, V., Crockett, M.J., Everett, J.A.C., Evgeniou, T., Gopnik, A., Jamison, J.C., Kim, T.W., Liao, S.M., Meyer, M.N., Mikhail, J., Opoku-Agyemang, K., Borg, J.S., Schroeder, J., Sinnott-Armstrong, W., Slavkovik, M., Tenenbaum, J.B.: Computational ethics. Trends Cognit. Sci. **26**(5), 388–405 (2022). https://doi.org/10.1016/j.tics.2022.02.009
4. Bronstein, J.M., Tagliati, M., Alterman, R.L., Lozano, A.M., Volkmann, J., Stefani, A., Horak, F.B., Okun, M.S., Foote, K.D., Krack, P., Pahwa, R., Henderson, J.M., Hariz, M.I., Bakay, R.A., Rezai, A., Marks, W.J., Jr., Moro, E., Vitek, J.L., Weaver, F.M., DeLong, M.R.: Deep brain stimulation for parkinson disease: an expert consensus and review of key issues. Arch. Neurol. **68**(2), 165 (2011). https://doi.org/10.1001/archneurol.2010.260
5. Carter, J.A., Clark, A., Palermos, S.O.: New humans?: ethics, trust, and the extended mind. Extended Epistemol. (2018). https://doi.org/10.1093/oso/9780198769811.003.0017
6. Clark, A.: Natural-born cyborgs: minds, technologies, and the future of human intelligence. Oxford University Press (2004). https://books.google.com/books?id=8JXaK3sREXQC. Accessed 17 Aug 2022
7. Clark, A.: Supersizing the mind: embodiment, action, and cognitive extension. Philos. Mind Ser. (2008). https://doi.org/10.1093/acprof:oso/9780195333213.001.0001
8. Clark, A., Chalmers, D.J.: The extended mind. Analysis **58**(1), 7 (1998)
9. Curchoe, C.L., Bormann, C.L.: Artificial intelligence and machine learning for human reproduction and embryology presented at ASRM and ESHRE 2018. J. Assist. Reprod. Genet. **36**(4), 591–600 (2019). https://doi.org/10.1007/s10815-019-01408-x
10. Deuschl, G., Schade-Brittinger, C., Krack, P., Volkmann, J., Schäfer, H., Bötzel, K., Daniels, C., Deutschländer, A., Dillmann, U., Eisner, W., Gruber, D., Hamel, W., Herzog, J., Hilker, R., Klebe, S., Kloß, M., Koy, J., Krause, M., Kupsch, A., Voges, J.: A randomized trial of deep-brain stimulation for Parkinson's disease. N. Engl. J. Med. **355**(9), 896–908 (2006). https://doi.org/10.1056/NEJMoa060281
11. Dilsizian, S.E., Siegel, E.L.: Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Curr. Cardiol. Rep. **16**(1), 1–8 (2014)
12. Fernandez, E.I., Ferreira, A.S., Cecílio, M.H.M., Chéles, D.S., de Souza, R.C.M., Nogueira, M.F.G., Rocha, J.C.: Artificial intelligence in the IVF laboratory: overview through the application of different types of algorithms for the classification of reproductive data. J. Assist. Reprod. Genet. **37**(10), 2359–2376 (2020). https://doi.org/10.1007/s10815-020-01881-9
13. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind. Mach. **28**(4), 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
14. Floridi, L., Cowls, J.: A unified framework of five principles for AI in society. Harvard Data Sci. Rev. **1**(1), 1–15 (2019). https://doi.org/10.1162/99608f92.8cd550d1
15. Ghassemi, M., Nsoesie, E.O.: In medicine, how do we machine learn anything real? Patterns (2022). https://doi.org/10.1016/j.patter.2021.100392
16. Goldenberg, S.L., Nir, G., Salcudean, S.E.: A new era: artificial intelligence and machine learning in prostate cancer. Nat. Rev. Urol. **16**(7), 391–403 (2019)
17. Gruetzemacher, R., Whittlestone, J.: The transformative potential of artificial intelligence. Futures **135**, 102884 (2022). https://doi.org/10.1016/j.futures.2021.102884
18. Hamet, P., Tremblay, J.: Artificial intelligence in medicine. Metabolism **69**, S36–S40 (2017)
19. Han, J., Davids, J., Ashrafian, H., Darzi, A., Elson, D.S., Sodergren, M.: A systematic review of robotic surgery: from supervised paradigms to fully autonomous robotic approaches. Int. J. Med. Robot. Comput. Assisted Surg. **18**(2), e2358 (2022)
20. Khan, M.S., Sanchez, F., Zhou, H.: 3-D printing of concrete: beyond horizons. Cement Concrete Res. **133**, 106070 (2020). https://doi.org/10.1016/j.cemconres.2020.106070
21. Mosavi, A., Salimi, M., FaizollahzadehArdabili, S., Rabczuk, T., Shamshirband, S., Varkonyi-Koczy, A.R.: State of the art of machine Learning models in energy systems, a systematic review. Energies (2019). https://doi.org/10.3390/en12071301
22. Nasiri, S., Khosravani, M.R.: Machine learning in predicting mechanical behavior of additively manufactured parts. J. Mater. Res. Technol. **14**, 1137–1153 (2021). https://doi.org/10.1016/j.jmrt.2021.07.004
23. Nicholas, P., Rossi, G., Williams, E., Bennett, M., Schork, T.: Integrating real-time multi-resolution scanning and machine learning for conformal robotic 3D printing in architecture. Int. J. Archit. Comput. **18**(4), 371–384 (2020). https://doi.org/10.1177/1478077120948203
24. Nussbaum, M.: Capabilities as fundamental entitlements: sen and social justice. Feminist Econ. **9**(2–3), 33 (2003)
25. Nussbaum, M.: Creating capabilities : the human development approach. Harvard University Press, Cambridge (2011)
26. Oosterlaken, I., & van den Hoven, J.: The capability approach, technology and design. Springer Netherlands. https://books.google.com/books?id=vO0DtXJaZvMC (2012).
27. Palermos, S.O.: Knowledge and cognitive integration. Synthese **191**(8), 1931–1951 (2014). https://doi.org/10.1007/s11229-013-0383-0
28. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J.: Carbon emissions and large neural network training. ArXiv Preprint ArXiv: 210410350. (2021)
29. Port, A. A., Kim, C., & Patel, M.: Deep sensory substitution: noninvasively enabling biological neural networks to receive input from artificial neural networks (2022). https://arxiv.org/abs/2005.13291. Accessed 17 Aug 2022
30. Pritchard, D.: Extended virtue epistemology. Inquiry **61**(5–6), 632–647 (2018). https://doi.org/10.1080/0020174X.2017.1355842
31. Rolnick, D., Donti, P.L., Kaack, L.H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A.S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A.: Tackling climate change with machine learning. ACM Comput. Surv. (CSUR) **55**(2), 1–96 (2022)
32. Rowlands, M.: The new science of the mind: from extended mind to embodied phenomenology. MIT PInress. https://doi.org/10.7551/mitpress/9780262014557.001.0001 (2010)
33. Sen, A.K.: Commodities and capabilities. Oxford University Press, India (1985a)

34. Sen, A. K.: Women, technology and sexual divisions. United Nations Conference on Trade and Development & United Nations International Research and Training Institute for the Advancement of Women. https://digitallibrary.un.org/record/83171/files/%5ETD_%5EUNCTAD_TT_79--UNCTAD_TT_79--TD_UNCTAD_TT_79-EN.pdf (1985b)
35. Sen, A.K.: Equality of what? In: Rawls, J., McMurrin, S.M. (eds.) Liberty, equality, and law: selected tanner lectures on moral philosophy. University of Utah Press, Salt Lake City (1987)
36. Strubell, E., Ganesh, A., & McCallum, A.: Energy and policy considerations for deep learning in NLP. ArXiv Preprint ArXiv:1906.02243. (2019)
37. Swain, J., VerMilyea, M.T., Meseguer, M., Ezcurra, D., Ezcurra, D., Letterie, G., Sánchez, P., Trew, G., Swain, J., Meseguer, M., Nayot, D., Campbell, A., Huangv, I., Choma, J., Loewke, K., Piqueras, M.P., Nader, P., Schindler, M., Lippolis, E., Group, F. A. I. F.: AI in the treatment of fertility: key considerations. J. Assisted Reprod. Genet. **37**(11), 2817–2824 (2020). https://doi.org/10.1007/s10815-020-01950-z
38. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R.: Intriguing properties of neural networks. ArXiv Preprint ArXiv:1312.6199. (2013)
39. Thiebes, S., Lins, S., Sunyaev, A.: Trustworthy artificial intelligence. Electron. Mark. **31**(2), 447–464 (2021). https://doi.org/10.1007/s12525-020-00441-4
40. Tóth, V., Parkkonen, L.: Autoencoding sensory substitution. https://doi.org/10.13140/RG.2.2.10576.87048 (2019)
41. Tuia, D., Kellenberger, B., Beery, S., Costelloe, B.R., Zuffi, S., Risse, B., Mathis, A., Mathis, M.W., van Langevelde, F., Burghardt, T., Kays, R., Klinck, H., Wikelski, M., Couzin, I.D., van Horn, G., Crofoot, M.C., Stewart, C.V., Berger-Wolf, T.: Perspectives in machine learning for wildlife conservation. Nat. Commun. **13**(1), 792 (2022). https://doi.org/10.1038/s41467-022-27980-y
42. Wiese, W., Friston, K.J.: AI ethics in computational psychiatry: From the neuroscience of consciousness to the ethics of consciousness. Behav. Brain Res. (2021). https://doi.org/10.1016/j.bbr.2021.113704