Ethics and Information Technology (2022) 24:4
https://doi.org/10.1007/s10676-022-09624-3

ORIGINAL PAPER

A sociotechnical perspective for the future of AI: narratives, inequalities, and human control

Laura Sartori[1] · Andreas Theodorou[2]

Accepted: 5 January 2022

Abstract

Different people have different perceptions about artificial intelligence (AI). It is extremely important to bring together all the alternative frames of thinking—from the various communities of developers, researchers, business leaders, policymakers, and citizens—to properly start acknowledging AI. This article highlights the 'fruitful collaboration' that sociology and AI could develop in both social and technical terms. We discuss how biases and unfairness are among the major challenges to be addressed in such a sociotechnical perspective. First, as intelligent machines reveal their nature of 'magnifying glasses' in the automation of existing inequalities, we show how the AI technical community is calling for transparency and explainability, accountability and contestability. Not to be considered as panaceas, they all contribute to ensuring human control in novel practices that include requirement, design and development methodologies for a fairer AI. Second, we elaborate on the mounting attention for technological narratives as technology is recognized as a social practice within a specific institutional context. Not only do narratives reflect organizing visions for society, but they also are a tangible sign of the traditional lines of social, economic, and political inequalities. We conclude with a call for a diverse approach within the AI community and a richer knowledge about narratives as they help in better addressing future technical developments, public debate, and policy. AI practice is interdisciplinary by nature and it will benefit from a socio-technical perspective.

Keywords Artificial intelligence · Sociology · Accountability · Transparency · Narratives · Inequalities · Human control

Introduction

Artificial intelligence (AI) is not a new field, it has just reached a new 'spring' after one of the many 'winters' (Boden, 2016; Floridi, 2020). As a matter of fact, we might be on the brink of a new winter since different actors (firms, individuals, media and institutions) have concretely started questioning the over-inflated expectations. It may be the multiple ongoing narratives, including the ones of moving from the traditional 'black-box approach' to the use of transparent and explainable methods (Guidotti, 2019a, 2019b), the 'scary'—but improbable—prospects of creating a 'superintelligence' that will convert humans into paperclips (Bostrom, 2014), or even of an 'AI race' between nations for the development of the 'ultimate' algorithm (Houser & Raymond, 2021). Then again, the term 'AI' means different things to different people; anything from data aggregation and manipulation to 'magic' (Theodorou & Dignum, 2020).

Yet, AI is neither a myth nor a threat to man (Samuel, 1962). In more actual terms, the welcome heated debate over AI for social good should not forget that AI systems are neither the ultimate chaos nor the ultimate panacea to social, political and economic issues of contemporary societies. Calls for responsible AI are mounting and eventually shedding light on many overseen social cutting issues. Responsible AI is built around three pillars: (i) governance; (ii) mechanisms; and (iii) means of participation (Dignum, 2019, 102–104). It is crucial to stop and think differently about our autonomous systems by considering them AI socio-technical systems, i.e. the combination of the technical component (i.e. the code and—if used—the data) and socio elements (i.e. the stakeholders responsible for the system

✉ Laura Sartori
l.sartori@unibo.it

Andreas Theodorou
andreas.theodorou@umu.se

[1] Department of Political and Social Sciences, University of Bologna, Via Bersaglieri 6, 40125 Bologna, Italy

[2] Department of Computer Science, Umeå University, 90187 Umeå, Sweden

Published online: 24 January 2022

and the society in which the system is deployed) (Dignum, 2019).

In the attempt to avoid futuristic oversimplification (Floridi, 2020), AI needs a multidisciplinary community. Afterall, when the field of Artificial Intelligence was first established at the now famous *Dartmouth Workshop*, albeit hosted at the mathematics departments, the participants included multiple psychologists, cognitive scientists, economists, and political scientists (McCarthy et al., 1955). Between the 'AI winters and springs', the focus of the field shifted to technical performance, ignoring the 'socio' part in the socio-technical systems. In recent years, sociology, among other social sciences, has looked back to the growing importance of AI with a scattered and ambivalent interest. Starting with big data analytics, it seemed sociology had to contribute mainly to the methodological sphere. As applications to social phenomena, all attention went on the newly developed tools in computational social sciences, such as agent-based models and the whole family of machine learning techniques (Bainbridge et al., 1994; Christin, 2020). Potential social impacts and implications received scant attention, although AI proved to have already entered the ways individuals, firms and public institutions organize processes of production, distribution, and exchange as well as consumption, public opinion, and politics. To encourage a socio-technical perspective to AI to flourish, our objective is to highlight some issues that are its building blocks.

First, we address the need for a useful sociology of AI. This claim has been advanced in the past just before the winter of AI. Now, before the mounting AI bubble bursts again, it urges to expand a sociological perspective that complements technical approaches to AI.

Second, this article focuses on the main challenges associated with autonomous systems, as their use spreads throughout our societies.

If much interest has been directed to technical possibilities and performances, we understand biases and unfairness as being among the major challenges to be addressed in such a sociotechnical perspective. As they are tightly related with real societal inequalities, we first notice advancements within the AI technical community that calls for transparency and explainability, accountability and contestability with some interesting insights from the social sciences. Then, a summary of the major results linking AI systems with inequalities is offered.

Third, the article aims at introducing the topic of narratives. Individuals, with their hopes and fears, do configure a technological imaginary that plays a crucial role for the spread, acceptance, and usage of any technology. How citizens—be they programmers, politicians, entrepreneurs or citizens—depict and account for AI may influence its own technical development. Not only do narratives reflect organizing visions for society, but they also are a tangible sign of the traditional lines of social, economic, and political inequalities. Knowledge about narratives helps in better addressing future technical developments, public debate, and policy.

In the following sections, we discuss each one of these issues in turn. Then, we wrap up our consideration for a socio-technical approach to AI linking it to the concept of AI for social good. What does 'good' mean for society? Far from the strict political philosophical considerations, it does not equate with 'fixing' social problems. Socially-aware AI systems are a first, not the ultimate, step into improving society and meeting societal challenges.

A sociology for AI

The interaction between AI and social sciences is a "fruitful arrangement." While social sciences contribute to the development of socio-beneficial systems—or even in the development of AI techniques—AI research has been contributing back to sociological research.

In parallel with other domains such as philosophy, sociology could contribute to a diverse understanding of AI. A sociological conception of AI adds up to the open line of investigation about wide social implications, offering crucial understanding for designers and developers trying to anticipate possible negative consequences. The need for a sociological conception of AI lies in the missing link to consistent social sciences empirical studies. Existing research is predominantly driven by the technical possibility (machine learning and neural network) applied to social and economic phenomena rather than being spurred by theoretically grounded research questions (Liu, 2021). To be more explicit, in the former case AI systems are more probably blind to social complexity (inequality, diversity, power structures) while in the latter sociological possibilities could inform and drive technological development. Here lies a strong potential for innovation in the current general discourse on AI. We need to elaborate on traditional questions such as how AI modelling can help develop social theory and methods as much as how social theoretical models might contribute to the development of AI.

In the past, following the ebbs and flows of AI developments in the computer science field, some sociologists tried to make a proper theoretical framework for a sociological conception of AI. As for other topics, sociology dealt with other disciplines' leftovers. It happened to money whose nature was exclusively investigated by economics while sociology was left to deal with residual aspects, such as values (Collins, 1979). History repeats when AI came along with a restricted view of sociological competence (Woolgar, 1985). Since Margaret Boden's work (1977), the notion of 'social' uncritically refers not to its nature (genesis) but to the uses of AI and their effects, restricting the scope of a sociological investigation. Yet, Woolgar underlines how distinctively social human behaviour is by having meaning-generating capabilities and how boundaries within social inquiry can be reasserted and pushed forward. More than systemic explanations based on strict formal models, interpretative sociological approaches can 'expand and elaborate' on that distinctiveness (Wolfe, 1991). Humans have minds that interpret the external reality, beyond the ability to follow instructions. With a 'mindful brain' (Edelman & Mountcastle, 1978) that software—based on algorithms—cannot have, they also make sense of reality with the aid of frames through which they organize social experience. Social frameworks provide implicit knowledge for understanding and moving in and out of different situations, ignoring inconsistencies and contradictions, as shown in innovative sociological work by ethnomethodologists Garfinkel (1967) and Goffman (1974). Not without critics, computer scientists—like Marvin Minsky as well as Robert Schank and Robert Abelson with their work on frames and knowledge-structures in the late Seventies—tried to organize the 'social life' for AI systems. Scripts, plans, and goals are shortcuts to understand a social situation, providing that implicit background necessary to 'understand' the 'social setting' for AI systems (Schwartz, 1989).

Far from fully explaining the entanglement of concepts such as society and institutions, agency, and intelligence, it is sufficient to say that there is a strong need for developing a sociology for AI. Its usefulness lies on the opportunities to argue both on the sociological origin as much as the social impacts of AI.

As we said at the beginning of this section, it is not just sociology contributing to AI, but also AI to sociology. Agent-based models, computer programs in which intelligent agents interact with each other in a set environment based on a defined set of rules, have been used to test social theories and examine macro-level effects from micro-level behaviour (Salganick, 2017). Examples of agent-based models used for sociological work include models on the appearance of modern human behaviour (Powell et al., 2009), costs of sharing information (Čače & Bryson, 2007), the evaluation of cooperation (Axelrod, 1997), political polarisation (Stewart et al., 2020) and multiple others. Other than the use of social simulations, AI techniques can be applied to provide new means of discovering and evaluating sociological findings (Molina & Garip, 2019).

Rising challenges

Between the ethics and governance of AI systems, much research interest has been devoted to their performance. However, there have been increasing calls in the development of the means, both in terms of technical and socio-legal solutions, to ensure the beneficial use of AI. We have identified three core challenges which we review in this section: (i) the opaque nature of machines; (ii) the guarantee of the respect for human agency and control of our autonomous artefacts; and (iii) the link to inequalities both as input to and output of AI systems.

To a comprehensive social and technical perspective to thrive, the biggest challenges AI faces are biases and discriminations, no novelties to human history. Intelligent machines replicate, duplicate, existing inequalities since they rely on biased dataset to start with. As magnifying glasses, they automate and amplify existing inequalities. Let alone legitimation (so far an underestimated dimension), the degree of discretion given to machines varies to the extent that they are often referred to as 'black boxes' operating with opacity.

Once more sociological research discovered a relationship between AI systems (and its broad constellation of related techniques) and inequalities, 'black boxedness,' and opacity. As a hot topic for all disciplines involved (from law to sociology and computer science), they started to be questioned, sustaining the demand for technical transparency and explainability. The quest for less biased AI systems came along as negative social implications were evident to the general public. Here, we first face what black box and opacity mean to transparency and explainability, then we turn to actual examples of automated inequalities to end with a scrutiny of the quest for AI for social good.

Black boxes and opacity in AI

Various AI—particularly machine learning—techniques have been making successes in producing accurate predictions and contributing to the ongoing 'AI hype'. However, many such systems remain opaque and obscure by communicating no understanding of the underlying real-time decision-making mechanisms (Burrell, 2016; Pasquale, 2015). Hence, people have been having trouble in understanding how an algorithm is built and why it produces a precise output (e.g. a decision on a loan, welfare service, college acceptance or job promotion). This is either due to design choice, e.g. economic or social factors, or technical limitations as some of the current most well-performing machine learning approaches do not lend themselves to explanations (Ananny & Crawford, 2018).

This black-box nature of intelligent systems makes interaction limited and uninformative for the end user. The lack of sufficient information regarding the emerging behaviour of a system results in its users creating inaccurate mental models (Wortham et al., 2017), which in turn may lead to them placing too much or too little trust in their system (Lee & See, 2004). Either case, this poses a safety risk as people may disuse or misuse the system and contribute to its credibility (Kemper & Kolkman, 2019). Furthermore, organisations have been exploiting the black-box nature of systems to deny the legal liability of their systems. We have already seen real-world examples of such practices, e.g. Apple's discriminatory credit card system (Nasiripour & Natarajan, 2019), Lufthansa's anti-monopoly (Bundeskartellamt, 2018), and others.

In response to these concerns, there has been an increasing push by academics and policymakers to make intelligent systems *transparent* and *explainable* (Barredo Arrieta et al., 2019; Guidotti et al., 2019a, 2019b; Miller, 2019; Theodorou et al., 2017). The former attribute implies that the system's decision-making mechanism is available for on-demand inspections of its status (Theodorou et al., 2017). Meanwhile, Explainable AI refers to the system being able to produce explanations of its behaviour, e.g. communicate the causes that caused an action (Miller, 2019; Pedreschi, 2019). While progress is being made in the technical challenges related to algorithmic transparency and explainability, there are still open-ended questions that require input from humanities and social sciences. Theodorou et al. (2017) state how transparency-related information should vary depending on the context that it is requested; different stakeholders, systems, and application domains all have different needs on the amount and way of presenting information. Failure to contextualise transparency information may lead to infobesity, i.e. overloading the user with information. Bringing back a sociological perspective into AI can help us not only to identify which information is relevant and meaningful for which stakeholder, but also when and how to display it.

In the attempt of opening and closely scrutinizing these black boxes, social sciences started to develop novel methodologies to debunk and penetrate these computational systems. Christin (2020) distinguishes three approaches: algorithmic audits, cultural and historical critiques, and ethnographic studies. The first relies on statistical analysis of online field experiments and a variety of quantitative datasets (from facial recognition systems to criminal records data). Although useful, it was criticized for looking into technical fixes and, thus, enacting the notion of black box (Abebe et al., 2020). Nevertheless, auditing algorithms is a first necessary step into questioning hierarchical classification systems, contributing to a greater fairness (D'Ignazio & Klein, 2020). The second approach is informed by a critical perspective that relates technological development of the reproduction of structural social and economic inequalities. While computational systems help to reduce society to metrics (Mau, 2018), the value of sorting and classification, commensuration, and standardization, is linked to wider societal processes such as globalization, surveillance and rationalization (Fourcade & Healy, 2017; Zuboff, 2019). A major critique relies on the scant attention to the micro level of analysis where individuals and institutions might shape the social construction and acceptance of computational systems. In tackling previous critiques, the third approach offers a novel view on how cultural imaginaries, institutional and organizational traditions are in play at the local level. Building on Social and Technology Studies, studies have focused especially on workers and how the computational turn changed their practices. Gig workers and normal users have gained increasing awareness of algorithms and AI systems, adapting their social practices and representations of on-demand and platform economy (Butcher, 2016; Rosenblat, 2017). The ethnographic approach has potential for showing how organizational culture responds when facing technology and how people use technology to counterbalance its negative social impacts (Elish & Boyd, 2017). Ethnography helps uncover the implicit and understand how data were cemented into the socio-technical system (Marda & Narayan, 2021).

Guaranteeing human control

However, we should not consider transparency and explainability as panaceas. The goal of transparency is to simply provide sufficient information to ensure attribution of accountability and, therefore, human control (Bryson & Theodorou, 2019). While the usage of approaches such as *human-in-the-loop* (Zanzotto, 2019) can provide human oversight, the concept of *human control* goes beyond technical oversight and instead includes the responsibility that lays in the development and deployment processes. In fact, socio-technical frameworks for *meaningful human control* aimed for high-risk situations, e.g. autonomous weapons, include the design and governance layers into what it means to have effective control (Horowitz & Scharre, 2015; Santoni de Sio & van den Hoven, 2018). Hence, human control can also be established through the use of well-established design practices. These practices may include requirements, design, and development methodologies, e.g. *Design for Value* (van den Hoven, 2005; Van de Poel, 2013) or the use of technical and ethical standards as means of demonstrating *due diligence* (Bryson & Winfield, 2018). Sociological knowledge not only enforces control, but it also offers insights about the existing structure of inequalities where social relations and structures, hierarchies, and organizations, are entangled. Incorporating this knowledge into the decision-making of key stakeholders and directly into a system's design, e.g. in its governance layer, can directly.

An important aspect of human control is our ability to effectively contest decisions made by a computer, i.e. *contestability*. Our ability to appeal to decisions made for us is considered a universal human right, while the GDPR makes explicit mention of algorithmic contestability (European

Parliament and Council of European Union, 2016). Contesting a decision requires much more than an explanation of the behaviour of the system; it requires us to understand and review a decision made by a system within the socio-legal context of where and when it was taken (Aler Tubella et al., 2020). Developing solutions to enable the contestability of systems we need to better understand the socio-legal norms that we want to verify our system's decision against. These norms are—as sociological research has shown again and again (Albright, 2019)—context dependent and are tightly linked to narratives, as we will see in the next section.

Countering this problem led academic and industrial research to 'explain' or even verify a system's decisions on the basis of specific social and ethical grounds: from societal acceptance to accountability, from individual/collective trust to fairness, from reducing sources of discrimination to equality enhancement. As Kasirzadeh (2021) points out "Because inherently complex and complicated AI ecosystems are connected with various stakeholders, institutions, values, and norms, the simplicity and locality of an explanation should not be the only virtue sought in the design of explainable AI", we need to widen the dominant—as for today—perspective on AI systems. This is a promising way to respond to challenges and contain risks. We need to consider the views of all these stakeholders, as their interpretation of values might be different from one to another (Aler Tubella et al., 2020). This variety of views is not unexpected, after all, there is no such thing as 'universal morality' (Turiel, 2002). Instead of trying a 'one-size-fits-all' definition for our ethical, social and cultural values, we should instead always try to make any definition—and claim compliance to said definition—explicit and transparent (Aler Tubella et al., 2019). While computer science can provide the means to formally represent the interpretations of values, sociology is needed to help us understand where and when an interpretation is valid and reliable.

Inequalities

Transparency and explainability are also tightly tied to the issue of *inequality*. Research has shown, there is an increasing 'algorithmic reliance' for individual's decisions in everyday life: from simple decisions, such as which movie to watch or restaurant to eat at (Paraschakis, 2017), to far more complex—and arguably important—choices about schools and universities. (O'Neil, 2016). Furthermore, we have seen the increasing use of intelligent systems to automate—and in some cases increase the amount of parameters taken into consideration—decisions with unequal socio-economic impact related to consumer credit (Aggarwal, 2020;); urban mobility (Rosenblat et al., 2017), courts (Larson et al., 2016), welfare (Eubanks, 2018); health (Obermeyer & Mullainathan, 2019) or territorial and logistic (Ingold & Soper, 2016) services.

Yet, if we look at two fundamental structural sources of inequalities, such as gender and race, we can see evidence on how AI systems are far from being neutral—let alone fair (Buolamwini & Gebru, 2018; Benjamin, 2019; Edelman et al., 2017; Hu, 2017; Kleinberg et al., 2019; Noble, 2018; Zhang et al., 2021). Although these systems may improve individual and social welfare (Taddeo & Floridi, 2018), they also have potential for creating a new social underclass (Benjamin, 2019) and digital poorhouses (Eubanks, 2018) through exclusionary intersectional practices (Park & Humphrey, 2019). Biases in AI do not exist in a vacuum nor are the product of algorithms. They exist in our own culture—including language—and are then obscured and propagated through the use of intelligent systems (Caliskan et al., 2017).

Several case studies are emblematic of non-neutral and discriminatory AI applications such as the Amazon hiring algorithm (Dastin, 2018) or the Apple credit scoring system (Methnani et al., 2021). The UK A-level grading fiasco represents an all-encompassing example of how an AI-system might be biased in both its inner design and training data.

Back in August 2020, the UK Office of Qualifications and Examinations Regulation overruled the results of A-level qualification (that certify school leaving and for university entrance) because of protests spurred around the country. To face the challenges brought up by the Pandemic, the UK Government decided—instead of moving the exams online or in a socially safe environment—to use an algorithm to award grades had the student taken the exam. Initially backed by a broad consensus among the public and policymakers, the use of such a grading algorithm soon revealed its unforeseen consequences. Students contended that the algorithm was biased: it made use of different sources of data that tended to underestimate teachers' individual assessments and to overemphasize the school's grading history. As a result, it penalized small schools with less stable distributions of grades or higher percentages of students with disadvantaged backgrounds (Clarke, 2020). Focusing on a predicted algorithm-based grade, the UK fiasco revealed how past inequalities could be reflected and automated into a score that requires a wider set of information to fairly assess educational achievements. The unforeseen social implications of an AI system operating in a complex ecosystem are now evident. To contain further fiascos, the failed UK experiment paves the way to a fruitful collaboration between technical practitioners and sociologists where technical (transparency, accountability and responsibility) and social (structural and educational inequalities) are jointly considered. As such, all stakeholders involved (in this case students, teachers, schools, ministry of education among others) are considered balancing needs for human control over AI systems and demands for equality in the society.

Overall, biases are unavoidable given the impossibility of data objectivity (D’Ignazio & Klein, 2020; Dignum, 2019; Leavy et al. 2020). After all, AI systems are an extension of our moral agency (Bryson, 2017; Theodorou, 2020) and the inequalities we naturally embody as social actors belonging to social, economic and political structures (e.g. social classes). As Caliskan et al. (2017) suggest, the use of AI does not only bring the risk of automating and magnifying inequalities, but it also offers the opportunity to use transparency in AI to better identify those biases in hope of their overcoming, or, at least, counterbalance with respect to society.

However, if we ever want to reach that stage, we need to build not only the tools, but also the culture for identifying, acknowledging, and addressing inequalities and discrimination in our societies. This culture could start by educating the AI community into understanding the social impact of its creations by linking mechanisms to ethics and values and part of social and technical relations (Theodorou & Dignum, 2020). It is about developers’ awareness of AI societal implications. Not only should there be proper training for developers, but also for all other stakeholders; including users and the general public that is indirectly affected by the technology. In the next section, we look at a broad overview of the current scenarios citizens usually juggle about.

Narratives and myths: organizing visions with concrete effects

Over the centuries, as any technology has developed, fictional and non-fictional narratives of fears and hopes came along. Associated with specific characteristics of a technology, the dominant narratives have had relevant and tangible effects. Narratives have the capacity to circulate and self-reinforce through patterns of repetition. As a specific representation of a technology, narratives also act as stories that brighten up mundane lives of individuals and societies offering a ‘technological myth’ (Mosco, 2004).

The way in which technologies are portrayed is crucial for their understanding and reception. Examples within the fields of domestic appliances (e.g. microwave) genetic modification or climate change offer interesting case studies for understanding how narratives could shape and influence further technological development and adoption as well as perception and confidence.

Speaking of technology, the nearest relevant example in time is the Internet. In the first decade of a widespread use, two influential alternative competing narratives stressed enthusiastic as much as catastrophic consequences on social relations (Kraut et al., 2002), sociality and trust (Norris, 2004; Wellman, 2001), work (Sproull & Kiesler, 1991), psychological well-being (Turkle, 1995) and democracy (Bimber, 2003). Further back, the telephone (Fischer, 1992) and the television (Bogart, 1956) shared a similar path when they first came around.

The same is happening to AI. While the effects on work (De Stefano, 2019; Frey & Osborne, 2017; GPAI, 2021) and democracy (Manheim & Kaplan, 2019; Schipper, 2020; Unver, 2018) have been among the first to catch the eye of social scientists, many on the wider social, economic, and political implications of AI are flourishing. As a result, extreme optimistic or pessimistic narratives could be pushed to their extremes, negatively impacting the future of AI. To this end it is important to keep in mind that, as previously said, AI is not new. For example, the reconstruction of the rise of AI from the 1950s and to early 1970s by the lens of a technological myth is telling about the underlying dynamics—among all the protagonists conveyed around a new emerging socio-technical system: from developers, to journalists, experts and users (Natale & Ballatore, 2020)—and their outputs. Despite what the narrative of an “AI winter” contended, it had continued to exercise a strong impact. As its technological myth went underground, the presence of a ‘socio-technical trajectory’ allowed for the creation of a community organized around a shared narrative of the future. It happened in the past. Messeri and Vertesi (2015) showed how the two unflown NASA missions in the Sixties were crucial to fastening together the planetary science community. Similarly, the current hype about AI is strongly organized around predicting and projecting the future, putting forward claims about potential uses and hypothetical performances. While reinforcing high expectation narratives around “an AI race between nations” or even of the development of a “superhuman Artificial General Intelligence,” the hype consolidates some narratives that produce concrete effects on how AI is developed, funded, used. As a matter of fact, narratives and myths act as ‘organizing visions’ (Dourish & Bell, 2011) embedded in specific cultural and institutional contexts. They also serve a performative role as they could obfuscate the real potentialities of technological advancements (Elish & Boyd, 2017).

This is why we turn deeper into AI narratives. Cave and Dihal (2019) synthesize the prevalent hopes and parallel fears in Western countries in the XX and XXI centuries. These dichotomies of hopes and fears help in disarticulating the complex imaginary made of fictional and non-fictional work about AI. Immortality, ease, gratification, and dominance are hopes that contrast fears like inhumanity, obsolescence, alienation and uprising. The former are positively transformational by nature as opposed to the inner instability brought about by the latter. As with any other technology, the first and foremost expectation is to provide healthy and longer lives while fighting against human decadence with an engineered and wired body. The second dichotomy ‘ease versus obsolescence’ put in contrast the human dream of being free from work and the concrete fear of becoming useless. If the dream can be traced back to the Iliad (around 800 BC), the latest fear is about automation and AI applied to industry and services, not to mention the Luddites revolt against the first industrial mechanical looms. AI systems could also gratify humans being present and attentive, friends and lovers. Yet, once technology mediates social relationships, a sense of alienation arises in both social and economic spheres, as classic sociologists such as Simmel and Marx famously argued in the late XIX century. The last dichotomy expands the concept of power as something that humankind has always strived for and that the society is filled with. AI is a tool that could help in gaining and maintaining a position of power (over other countries or specific groups of people), whereas as an agent it could take control over humans.

As context dependent as they might be, narratives are nonetheless a useful instrument to decipher the cultural and institutional frameworks in which they are rooted, produced and consumed. As previously mentioned, narratives, myths and AI-technology are entangled in such a way that understanding underlining hopes and fears might support or avoid a specific future development, public reception, and regulation. Hopes and fears stand cheek to cheek by jowl of technical advancements.

Implications of the dominant narratives

Technologies are embedded within larger social systems and processes, inscribed with the rules, values and interests of a typically dominant group. So do narratives. They reveal important glimpses on the existing structure of inequalities that, through an increasing reliance on formalized algorithmic and AI techniques, could be automated. Demystifying dominant narratives helps, for example, in breaking this path for automation through the reproduction of stereotypes. That is to say, it breaks the traditional loop that sees a dominant group building its own narrative rooted in ordinary, but often overlooked, class, gender and racial inequalities.

Through a gendered reading of AI and automata, Adams (2020) recently contributed to expose some of the self-reinforcing power asymmetries reflected in the dominant narratives. Hopes and fears offer insights on the socio-cultural schemes about the role of women within different societies. At the same time, in reproducing dominant schemes, they contribute to shaping current realities. The choice for feminine voices and names of personal assistants like Siri and Alexa is enlightening. We want technology to help us while we keep the control-seat. So far, the solution has lied in turning to prevalent stereotypes that shape women as reassuring dedicated subjects for caring (Lerner, 2018) with a 'total availability' (Cross, 2016). Not even a chance 'to blush' (Unesco, 2019).

Again on gender, research shows biases in recruitment because the dedicated AI system was trained on decades-old datasets where fewer women were applying or promoted (e.g. Amazon HR in Dastin, 2018). While biases in facial recognition systems are unable to detect black faces (Buolamwini & Gebru, 2018), language translation reflects routine activities as gendered (Lee, 2018; Marcelo et al., 2020). Google Translate initially did not offer gender-specific translation for languages bearing neutral, like Hungarian or Turkish. All professions were referred to as male while domestic work was attributed to women. Now, this is addressed in single-sentence translations by showing the translation for both genders. Albeit an improvement, the use of neutral pronouns could promote further inclusivity. At the time of writing, in multi-sentence paragraphs, the gender-specific translations still occur. In this direction, addressing some of the most well-known biases in the AI field promises to bring about positive effects for research on societal impacts and implications. Also, a more positive and engaging discussion could benefit people's knowledge, perception, and reception of technological innovation in their daily life.

Previous research highlighted at least three directions for further reflection. First, there is a disconnect between narratives and actual science. So far, public knowledge on what AI consists of and what its implications are is poor (Royal Society, 2017; Zhang & Zafoe, 2019). Among others, one reason is found in the relation between technology and magic. In a self-reinforcing circle, the former promises technical efficiency while the latter tells an idealized version of technology. Both sustain a rich imaginary that tends to depart from reality.

Second, tales of fear dominate over more positive scenarios (Cave & Dihal, 2019; Royal Society, 2019). Although larger quotas of the population are acknowledging awareness of the potential benefits of AI, negative consequences catch the eye in the scientific and public debates. At first, talking about the future of work led to a pessimistic scenario where robots were to steal people's jobs within a short period of time (Frey and Osborne revised in 2017 their previous gloomier predictions). One of the first surveys targeting Machine Learning showed that it evoked standard ideas associated with increased efficiency, accuracy, objectivity. Yet, concerns about potential physical harms, job substitution or limiting individual choices were prevalent (Royal Society, 2017).

Third, plenty of evidence highlights a lack of diversity on both the production and consumption sides of AI (Broussard, 2018; Zhang et al, 2021). Diversity is missing in the development of the technology as much as in the design of AI applications. While the latter are naturally prone to reflect and reproduce biases and inequalities, lack of ex-ante diversity is not exempted from mimicking inequalities. Data too needs to be diversified (D'Ignazio & Klein, 2020; Eubanks,

2018) as limited and poor data lead to reinforcing social and economic existing biases (Crawford et al., 2016).

The perceived and concrete take that individuals have on the technology also influences the rise of narratives centered on responsibility. With regard to AI, discussions on abstract or long-term effects could lead to narratives detached from daily life where individuals feel disengaged and not responsible. Moreover, investigating narratives strongly involves ethics (Ward, 2006) because the way we pursue and reproduce our myths and organizing visions has consequences. There lies the need for a truly human-centered AI that could support honest narratives for a better development, public reception and regulation. Whether developers, citizens or policy makers, more realistic narratives help into projecting the future of AI.

Overall, tackling ignorance and false fears is a way to better address future technical developments, correctly direct public knowledge, debate, and policy in a fruitful collaboration, bidirectional relation, between sociology and AI.

Conclusions and future work

As we have been emphasising in our paper, AI practice is interdisciplinary by nature and it will benefit from a different take on technical interventions. Nor superior nor more appropriate, technical considerations (such as objectivity, fairness, and accuracy) should go in parallel with other types of knowledge useful for social change (Green, 2019). What issues to face, what data to use and what solutions to implement are compelling, not old-fashioned, questions.

It is not always a question of efficiency and accuracy, but also it is about inclusivity by bridging the gap between technical and social research in AI. In addition to responsibility, AI should be inclusive, built upon quality data that comprises gender, education, ethnicity, and all of the other social and economic differences that are sometimes determining factors for inequality. Quality data not only means to make it respectful of privacy but to make it inclusive when it comes to social concerns and purposes. Data allows for model and algorithms development that need, in turn, to be closely oversighted through mandatory requirements (Dignum et al., 2020). From here, the spread of praises for a 'New Deal' on data and its diversification and democratization (Benjamin, 2019) while pre-emptive and independent algorithmic impacts assessment tools are important elements to consider for regulation.

Once both positive and negative technical potentialities are considered, there is a chance to overcome optimistic or pessimistic scenarios linked to the major narratives discussed earlier. Not only is there a chance to tackle technological potential flaws, but also the current state of the art allows for a fresh start and a new narrative about technology. As argued more than twenty years ago (Suchman et al., 1999), technologies are social practices that can be assessed within a specific social and institutional context. As such, practices with social and environmental impacts (Vinuesa et al., 2020) have different outcomes and impacts on communities of citizens and workers. There is a potential for democratization of the internal processes to the AI communities, but also for equalization of the unbalanced outputs of current AI systems.

Regardless of the potential benefits, we might be at the brink of a new AI winter. The overhype around AI ignores that AI systems currently available are usable not in a faraway tech laboratory, but in our daily routines at home and at work. Hopes and fears about benefits and dangers play a central role in the people, businesses, and public bodies' adoption. At the same time, to effectively change and open up the dialog around these narratives, a call for diversity on the production side is necessary as well. Different people think differently about AI. It is extremely important to bring together alternative frames of thinking—in the community of developers, business, and citizens—to properly start reflecting about AI as socio-technical systems in both social and technical terms.

The relevance of technological myths and narratives applies to AI as much as previous technologies, especially, in the current hype bubble. Not only is this relevant for developers, but also for citizens and politicians. People's awareness and trust are crucial for adoption and usage of new technologies, like it has been for the telephone or the television. As historian Kranzberg (1986) noted, "technology is neither good nor bad, nor is it neutral". Technological determinism offers an easy trap to fall into when only developers participate in the design and construction of AI technology. Likewise, path dependence and lock-ins tricks easily fall within the comfort zone of programmers and policy makers. If we consider AI as a sociotechnical system, we are to include all participants in the process of construction in a co-creation approach. Hence, research could also shed light on the legitimization mechanisms underlying the relationship between social and artificial agents. Since technology is not any magic, the relevance of narratives in shaping current realities is a strong call for citizens—with their perceptions and beliefs—to sit at the table for the future of AI.

To address this lack of knowledge, we are currently working on a comparative project on Italian and Swedish narratives. Our research project, conducted under the umbrella of HumanE-AI-Net, aims to provide us with a better understanding of how narratives related to AI are formed and their relationship with the perceived trust, hopes, and fears in the technology.

Acknowledgements The authors would like to thank *HumanE-AI-Net* project, which has received funds from the European Union's Horizon

2020 research and innovation programme under Grant agreement 952026. Theodorou was also supported by the Knut Och Alice Wallenberg Foundation under Grant 2020:0221.

References

Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., and Robinson, D.G. (2020). Roles for computing in social computing in social change. In: Conference on Fairness, Accountability, and Transparency (FAT* '20)

Adams, R. (2020). *Helen A'Loy* and other tales of female automata: A gendered reading of the narratives of hopes and fears of intelligent machines and artificial intelligence. *AI &amp; Society, 35*, 569–579. https://doi.org/10.1007/s00146-019-00918-7

Aggarwal, N. (2020). The norms of algorithmic credit scoring. *Cambridge Law Journal*. https://doi.org/10.2139/ssrn.3569083

Albright, B. (2019). If you give a judge a risk score: Evidence from Kentucky bail decisions. Retrieved from https://thelittledataset.com/about_files/albright_judge_score.pdf

Aler Tubella, A., Theodorou, A., Dignum, F., and Dignum, V. (2019). Governance by Glass-Box: Implementing Transparent Moral Bounds for AI Behaviour. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI). DOI: https://doi.org/10.24963/ijcai.2019/802

AlerTubella, A., Theodorou, A., Dignum, V., & Michael, L. (2020). Contestable black boxes. *RuleML+RR*. Springer.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media &amp; Society*. https://doi.org/10.1177/1461444816676645

Axelrod, R. (1997). *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton University Press.

Bainbridge, W. S., Brent, E. E., Carley, K. M., Heise, D. R., Macy, M. W., Markovsky, B., & Skvoretz, J. (1994). Artificial social intelligence. *Annual Review of Sociology, 20*(1), 407–436.

BarredoArrieta, A., Diaz Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado González, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, V. R., Chatila, R., & Herrera, F. (2019). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*. https://doi.org/10.1016/j.inffus.2019.12.012

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Wiley.

Bimber, B. (2003). *Information and American democracy*. Cambridge University Press.

Boden, M. (1977). *Artificial intelligence and natural man*. MIT Press.

Boden, M. (2016). *AI: Its nature and future*. Oxford University Press.

Bogart, L. (1956). *The age of television: A study of viewing habits and the impact of television on American life*. Ungar Pub Co.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies* (1st ed.). Oxford University Press Inc.

Broussard, M. (2018). *Artificial unintelligence*. MIT Press.

Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. *Artificial Intelligence Law, 25*, 273–291. https://doi.org/10.1007/s10506-017-9214-9

Bryson, J. J., & Theodorou, A. (2019). How society can maintain human-centric artificial intelligence. In M. Toivonen-Noro, E. Saari, H. Melkas, & M. Hasu (Eds.), *Human-centered digitalization and services* (pp. 305–323). Springer.

Bryson, J. J., & Winfield, A. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer, 50*(5), 116–119. https://doi.org/10.1109/MC.2017.154

Bucher, T. (2016). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication &amp; Society, 20*(1), 30–44.

Bundeskartellamt. (2018). No proceeding against Lufthansa for abusive pricing. Retrieved from https://www.bundeskartellamt.de/SharedDocs/Entscheidung/EN/Fallberichte/Missbrauchsaufsicht/2018/B9-175-17.pdf?__blob=publicationFile&v=2

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency (FAT*), No. 81. pp. 77–91

Burrell, J. (2016). How the machine "Thinks": Understanding opacity in machine learning algorithms. *Big Data &amp; Society*. https://doi.org/10.1177/2053951715622512

Čače, I., & Bryson, J. J. (2007). Agent based modelling of communication costs: Why information can be free. In C. Lyon, C. L. Nehaniv, & A. Cangelosi (Eds.), *Emergence of communication and language*. Springer.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Cave, S., & Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence, 1*, 74–78.

Christin, A. (2020). The ethnographer and the algorithm: Beyond the black box. *Theory &amp; Society, 49*, 897–918. https://doi.org/10.1007/s11186-020-09411-3

Clarke, M. (2020) Examinations and high stakes decision making in the era of COVID-19. Retrieved from https://blogs.worldbank.org/education/examinations-and-high-stakes-decision-making-era-covid-19

Collins, R. (1979). The bankers by Martin Mayer. *American Journal of Sociology, 85*(1), 190–194.

Crawford, K., Whittaker, M., Elish, M.C., Barocas, S., Plasek, A., Ferryman, K. (2016). The AI now report: The social and economic implications of artificial intelligence technologies in the near-term. Report prepared for the AI now public symposium, hosted by the White House and New York University's Information Law Institute. Retrieved from https://artificialintelligencenow.com/media/documents/AINowSummaryReport_3.pdf

Cross, K (2016). When robots are an instrument of male desire. Retrieved from https://medium.com/theestablishment/when-robots-are-an-instrument-of-male-desire-ad1567575a3d.

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press.

Dastin, J. (2018). Amazon scrapped a secret AI recruitment tool that showed bias against women. Reuters 10 October 2018

De Stefano, V. (2019). Introduction: Automation, artificial intelligence, and labour protection. *Comparative Labor Law &amp; Policy Journal, 41*, 15.

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Switzerland: Springer Nature. https://doi.org/10.1007/978-3-030-30371-6.

Dignum, V., Muller, C., and Theodorou, A. (2020). Final analysis of the EU whitepaper on AI, June 12th, ALLAI

Dourish, P., & Bell, G. (2011). *Divining a digital future: Mess and mythology in ubiquitous computing*. The MIT Press.

Edelman, B. L., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics, 9*(2), 1–22.

Edelman, G. M., & Mountcastle, V. B. (1978). *The mindful brain: Cortical organization and the group-selective theory of higher brain function*. MIT Press.

Elish, M. C., & Boyd, D. (2017). Situating methods in the magic of big data and artificial intelligence. *Communication Monographs, 85*(1), 57–80.

Eubanks, V. (2018). *Automating inequality. How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

European Parliament and Council of European Union (2016) General data protection regulations (GDPR). Pub. L. No. 2016/679

Fischer, C. (1992). *America calling*. University of California Press.

Floridi, L. (2020). AI and its new winter: From myths to realities. *Philosophy &amp; Technology*. https://doi.org/10.1007/s13347-020-00396-6

Fourcade, M., & Healy, K. (2017). Seeing like a market. *Socio-Economic Review, 15*(1), 9–29.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technology, Forecasting and Social Change, 114*, 254–280.

Garfinkel, H. (1967). *Studies in ethnomethodology*. Prentice-Hall.

Goffman, E. (1974). *Frame analysis*. Harvard University Press.

GPAI (2021). Working group on the future of work. Retrieved from https://gpai.ai/projects/future-of-work/

Green, B. (2019). "Good" isn't enough. AI for social good workshop (NeurIPS2019)

Guidotti, R., Monreale, A., & Pedreschi, D. (2019a). The AI black box explanation problem. *ERCIM News, 116*, 12–13.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019b). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 93.

Horowitz, M., & Scharre, P. (2015). Meaningful human control in weapon systems: A primer, Working paper (Center for a New American Security).

Houser, K., & Raymond, A. (2020). It is time to move beyond the 'AI Race' narrative: Why investment and international cooperation must win the day. *Northwestern Journal of Technology and Intellectual Property, 18*, 129.

Hu, M. (2017). Algorithmic Jim Crow. *Fordham Law Review, 86*, 633.

Ingold, D., and Soper, S. (2016). Amazon doesn't consider the race of its customers. Should it?. Bloomberg. Retrieved https://www.bloomberg.com/graphics/2016-amazon-same-day/

Kasirzadeh, A. (2021). Reasons, values, stakeholders: A philosophical framework for explainable artificial intelligence. In: Conference on Fairness, Accountability, and Transparency (FAccT '21). DOI:https://doi.org/10.1145/3442188.3445866

Kemper, D., & Kolkman. (2019). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society, 22*(14), 2081–2096.

Kleinberg, J., Ludwig, J., Mullainathan, S., Sunstein, C.R. (2019). Discrimination in the age of algorithms. National Bureau of Economic Research

Kranzberg, M. (1986). Technology and history: Kranzberg's laws. *Technology and Culture, 27*(3), 544–560.

Kraut, R., Kiesler, S., Boneva, B., Cummings, J., Helgeson, V., & Crawford, A. (2002). Internet paradox revisited. *Journal of Social Issues, 58*(1), 49–74.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). *How we analyzed the COMPAS recidivism algorithm*. Propublica.

Leavy, S. O'Sullivan, B. and Siapera, E. (2020). Data, power and bias in artificial intelligence. Retrieved from https://arxiv.org/abs/2008.07341

Lee, D. (2018). *Google translate now offers gender-specific translations for some languages*. The Verge.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lerner, S. (2018). NHS might replace nurses with robot medics such as carebots: Could this be the future of medicine? Tech Times. https://www.techtimes.com/articles/229952/20180611/nhs-might-replace-nurses-with-robot-medics-such-as-carebots-could-this-be-the-future-of-medicine.htm.

Liu, Z. (2021). Sociological perspectives on artificial intelligence: A typological reading. *Sociology Compass, 15*(3), e12851.

Manheim, K. M., & Kaplan, L. (2019). Artificial intelligence: Risks to privacy and democracy. *Yale Journal of Law and Technology, 21*, 106.

Marcelo, O. R., Prates, P. H., Avelar, L., & Lamb, C. (2020). Assessing gender bias in machine translation: A case study with Google translate. *Neural Computing and Applications, 32*, 6363–6381. https://doi.org/10.1007/s00521-019-04144-6

Marda, V., & Narayan, S. (2021). On the importance of ethnographic methods in AI research. *Nature Machine Intelligence, 2*(3), 187–189.

Mau, S. (2019). *The metric society: On the quantification of the social*. Wiley.

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. E. (1955). A proposal for the dartmouth summer research project on artificial intelligence. *AI Magazine, 27*, 12.

Messeri, L., & Vertesi, J. (2015). The greatest missions never flown: Anticipatory discourse and the "Projectory" in technological communities. *Technology and Culture, 56*(1), 54–85.

Methnani, L., AlerTubella, A., Dignun, V., & Theodorou, A. (2021). Let me take over: Variable autonomy for meaningful human control. *Frontiers in AI*. https://doi.org/10.3389/frai.2021.737072

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.

Molina, M., & Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology, 45*(1), 27–45.

Mosco, V. (2004). *The digital sublime*. MIT Press.

Nasiripour, S., Natarajan, S. (2019). Apple co-founder says Goldman's apple card algorithm discriminates. Bloomberg. Retrieved from https://www.bloomberg.com/news/articles/2019-11-10/apple-co-founder-says-goldman-s-apple-card-algo-discriminates

Natale, S., & Ballatore, A. (2020). Imagining the thinking machine: Technological myths and the rise of artificial intelligence. *Convergence, 26*(1), 3–18.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Norris, P. (2004). The bridging and bonding role of online communities. In P. Howard & S. Jones (Eds.), *Society online*. Sage.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Books.

Paraschakis, D. (2017). Towards an ethical recommendation framework. In: 11th International Conference on Research Challenges in Information Science (RCIS). DOI: https://doi.org/10.1109/RCIS.2017.7956539.

Park, S., & Humphry, J. (2019). Exclusion by design: Intersections of social, digital and data exclusion. *Information, Communication*

&amp; Society, 22(7), 934–953. https://doi.org/10.1080/1369118X.2019.1606266

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, F., & Turini, F. (2019). Meaningful explanations of black box AI decision systems. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*, 9780–9784.

Powell, A., Shennan, S., & Thomas, M. G. (2009). Late Pleistocene demography and the appearance of modern human behavior. *Science, 324*(5932), 1298–1301. https://doi.org/10.1126/science.1170165

Rosenblat, A., Levy, K., Barocas, S., & Hwang, T. (2017). Discriminating tastes: Uber's customer ratings as vehicles for workplace discrimination. *Policy &amp; Internet, 9*(3), 256–279.

Royal Society. (2017). *Machine learning: The power and promise of computers that learn by example*. The Royal Society.

Royal Society. (2018). *Portrayals and perceptions of AI and why they matter*. The Royal Society.

Salganick, M. (2017). *Bit by bit: Social research in the digital age*. Princeton University Press.

Samuel, A. L. (1962). Artificial intelligence: A frontier of automation. *The ANNALS of the American Academy of Political and Social Science, 340*(1), 10–20. https://doi.org/10.1177/000271626234000103

Santoni de Sio, F., & van den Hoven J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Front Robot AI, 5*, 5. https://doi.org/10.3389/frobt.2018.00015.

Schippers, B. (2020). Artificial intelligence and democratic politics. *Political Insight, 11*(1), 32–35. https://doi.org/10.1177/2041905820911746

Schwartz, R. D. (1989). Artificial intelligence as a sociological phenomenon. *The Canadian Journal of Sociology / Cahiers Canadiens de Sociologie, 14*(2), 179–202. https://doi.org/10.2307/3341290.

Sproull, L., & Kiesler, S. (1991). *Connections. New ways of working in the networked organization*. MIT Press.

Stewart, A. J., McCarty, N., & Bryson, J. J. (2020). Polarization under rising inequality and economic decline. *Science Advances*. https://doi.org/10.1126/sciadv.abd4201

Suchman, L., Blomberg, J., Orr, J. E., & Trigg, R. (1999). Reconstructing technologies as social practice. *American Behavioral Scientist, 43*(3), 392–408. https://doi.org/10.1177/00027649921955335

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751–752.

Theodorou, A. (2020). Why artificial intelligence is a matter of design. In B. P. Goecke & A. M. der Pütten (Eds.), *Artificial intelligence* (pp. 105–131). Brill and Mentis.

Theodorou, A., & Dignum, V. (2020). Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence, 2*(1), 10–12. https://doi.org/10.1038/s42256-019-0136-y

Theodorou, A., Wortham, R. H., & Bryson, J. J. (2017). Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science, 29*(3), 230–241. https://doi.org/10.1080/09540091.2017.1310182

Turiel, E. (2002). *The culture of morality: Social development, context, and conflict*. Cambridge University Press.

Turkle, S. (1995). *Life on the screen: Identity in the age of the internet*. Weidenfeld & Nicolson.

UNESCO (2019). I'd blush if I could: Closing gender divides in digital skills through education. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000367416

Ünver, H. A. (2018). *Artificial intelligence, authoritarianism and the future of political systems*. Centre for Economics and Foreign Policy Studies.

Van de Poel, I. (2013). Translating values into design requirements. *Philosophy and engineering: Reflections on practice, principles and process* (pp. 253–266). Springer.

van den Hoven, J. (2005). Design for values and values for design. *Journal of the Australian Computer Society, 7*(2), 4–7.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*. https://doi.org/10.1038/s41467-019-14108-y

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics*. https://doi.org/10.1126/scirobotics.aan6080

Ward, G. (2006). Narrative and ethics: The structures of believing and the practices of hope. *Literature and Theology, 20*(4), 438–461.

Wellman, B., Haase, A. Q., Witte, J., & Hampton, K. (2001). Does the internet increase, decrease, or supplement social capital?: Social networks, participation, and community commitment. *American Behavioral Scientist, 45*(3), 436–455. https://doi.org/10.1177/00027640121957286

Wolfe, A. (1991). Mind, Self, Society, and Computer: Artificial Intelligence and the Sociology of Mind. *American Journal of Sociology, 96*(5), 1073–1096.

Woolgar, S. (1985). Why not a sociology of machines? The case of sociology and artificial intelligence. *Sociology, 19*, 557–572.

Wortham, R. H., Theodorou, A., & Bryson, J. J. (2017). Robot transparency: Improving understanding of intelligent behaviour for designers and users. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 274–289). Springer.

Zanzotto, M. F. (2019). Viewpoint: Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research, 64*(1), 243–252. https://doi.org/10.1613/jair.1.11345

Zhang, B., & Dafoe, A. (2019). *Artificial intelligence: American attitudes and trends*. Future of Humanity Institute, University of Oxford.

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., Shoham, M., Clark, J., & Perrault, R. (2021). *The AI index 2021 annual report*. Human-Centered AI Institute, Stanford University.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Public Affairs.