Toward accountable human-centered AI: rationale and promising directions

Junaid Qadir
Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar and Department of Electrical Engineering, Information Technology University, Lahore, Pakistan

Mohammad Qamar Islam
Department of Electrical Engineering, Information Technology University, Lahore, Pakistan, and

Ala Al-Fuqaha
Information and Computing Technology (ICT) Division, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

Received 9 June 2021
Revised 14 November 2021
Accepted 14 December 2021

Abstract

Purpose – Along with the various beneficial uses of artificial intelligence (AI), there are various unsavory concomitants including the inscrutability of AI tools (and the opaqueness of their mechanisms), the fragility of AI models under adversarial settings, the vulnerability of AI models to bias throughout their pipeline, the high planetary cost of running large AI models and the emergence of exploitative surveillance capitalism-based economic logic built on AI technology. This study aims to document these harms of AI technology and study how these technologies and their developers and users can be made more accountable.

Design/methodology/approach – Due to the nature of the problem, a holistic, multi-pronged approach is required to understand and counter these potential harms. This paper identifies the rationale for urgently focusing on human-centered AI and provide an outlook of promising directions including technical proposals.

Findings – AI has the potential to benefit the entire society, but there remains an increased risk for vulnerable segments of society. This paper provides a general survey of the various approaches proposed in the literature to make AI technology more accountable. This paper reports that the development of ethical accountable AI design requires the confluence and collaboration of many fields (ethical, philosophical, legal, political and technical) and that lack of diversity is a problem plaguing the state of the art in AI.

Originality/value – This paper provides a timely synthesis of the various technosocial proposals in the literature spanning technical areas such as interpretable and explainable AI; algorithmic auditability; as well as policy-making challenges and efforts that can operationalize ethical AI and help in making AI accountable. This paper also identifies and shares promising future directions of research.

Keywords Accountable Artificial intelligence, Human-centered AI, AI ethics

Paper type General review

This publication was made possible by NPRP Grant# [13S-0206–200273] from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

The first author would like to acknowledge Adnan Qayyum for his help in drawing the figures in this paper.

Journal of Information, Communication and Ethics in Society
Vol. 20 No. 2, 2022
pp. 329-342

1477-996X
DOI 10.1108/JICES-06-2021-0059

1. Introduction

While artificial intelligence (AI) technology offers various conveniences and opportunities for doing social good (Ali *et al.*, 2016), AI technology is not without harmful concomitants. AI can both promote and inhibit human development (Latif *et al.*, 2019; Vinuesa *et al.*, 2020). The uncritical adoption of AI technology, and its use without due processes and governance frameworks, can unleash massive damage ranging from engendering inequality, disadvantaging minorities and destabilizing the basic fabric of society (Buolamwini and Gebru, 2018; Helbing, 2019). In the words of author Cathy O'Neil, these technologies act as "weapons of math destruction" (O'Neil, 2016).

AI technology is also entwined with ethical and moral dilemmas that poses several threats to human and social flourishing (Helbing *et al.*, 2019; Denning and Denning, 2020). There are significant risks of unethical use of large-scale collected data that drives modern AI. The data itself may be unrepresentative or biased. The AI pipeline is also beset with privacy and security challenges with the potential of adversarial attacks. In addition, AI technology runs on top of a surveillance capitalism ecosystem with well-entrenched prioritization to profit-making rather than any real commitment to serving human interests (Zuboff, 2019). Therefore, although there is not significant interest in developing ethical AI, transforming this intention and desire into practice has been very difficult due to the challenges involved in operationalizing ethical AI arising from various vested interests whose interest is in the status quo being maintained so that AI could be used for whatever that leads to more profit.

It is also worth exploring the philosophy of technology and how it shapes how people perceive technology and think about responsible use of technology. Over the past century and a half, the field of philosophy of technology can be collected into two competing approaches: technological determinism and technological instrumentalism (Newport, 2020). According to *technological instrumentalism*, technologies are neutral and instrumental. Humans are therefore in the driving seat regardless of how the technology is designed. The focus is solely on human behaviors and contexts. *Technological determinism*, on the other hand, posits that features of technology tools can drive human behavior in unexpected directions. Recently, experts are beginning to challenge the validity of the commonly accepted instrumentalist philosophy, as it is ill-suited to tackle some of the more complicated questions at this time of rapid technological development (Newport, 2020). The perspective of technological determinism also brings engineers into the picture and holds them responsible for the outcomes of their products. Here, this becomes another measure of performance to measure and improve.

Our purpose with this paper is to explore the impact of AI on society more holistically and to propose a way forward for AI, so that it becomes accountable ethical and human-centered. To undertake this study, we engage in a broad review of literature and highlight a taxonomy of ethical challenges that confront the use of AI. Thereafter, we highlight how we can steer AI so that we can reap its benefits and promote human development but without suffering from its harms that inhibit human well-being or enhances inequality. Put differently, we highlight the current risks posed by AI and provide an overview of promising directions that can be used to create more accountable human-centered AI.

Our main contribution in this paper is that *firstly*, we comprehensively review the ethical and social challenges related to the practice of AI, and *secondly*, we discuss how we can solve these problems with technical solutions (such as algorithmic auditability, AI explanability) using some promising technical and non-technical directions.

The remainder of the paper is organized in the following way. The potential downsides of AI, and the ethical challenges associated with AI, are introduced in Section 2. The

discovery of these harms and ethical challenges has created an AI ethics bandwagon. The researchers in this area have made progress but the field overall has remained toothless and not yet fully effective. The strengths and limitations of works aiming at ethical AI are described in Section 3. We will discuss how we can get out of this impasse by identifying promising directions that can help operationalize accountable human-centered AI in Section 4. Finally, the paper is concluded in Section 5.

2. Caveat emptor: let the buyer of artificial intelligence be aware

A lot of research informs us that AI technology is a double-edged sword in that it can both promote and inhibit human development (Latif *et al.*, 2019; Vinuesa *et al.*, 2020). There are various concerns related to modern machine learning techniques including bias, lack of robustness, lack of transparency and the high cost of training large AI models. We find substantial reports and evidence in the literature and practice that indicates that AI technology, along with its various beneficial uses, has various unsavory concomitants as noted next and illustrated in Figure 1:

- The opaqueness of their mechanisms (O'Neil, 2016; Lipton, 2018).
- The inscrutability of AI tools and their lack of accountability (Raji *et al.*, 2020).
- The fragility of AI models under adversarial settings (Marcus and Davis, 2019)
- The vulnerability of AI models to bias throughout their pipeline (Suresh and Guttag, 2019)
- The high planetary cost of running large AI models (Crawford, 2021).
- The emergence of exploitative surveillance capitalism-based economic logic built on AI technology (Zuboff, 2019).

The various harms of AI-driven modern technology are documented in the "ledger of harms" [1] curated by the Center for Humane Technology – co-founded by Tristan Harris, a prominent technology critic, and formerly a Google engineer. The ledger includes the causation of disruption to social relationships (less empathy, more confusion), physical and

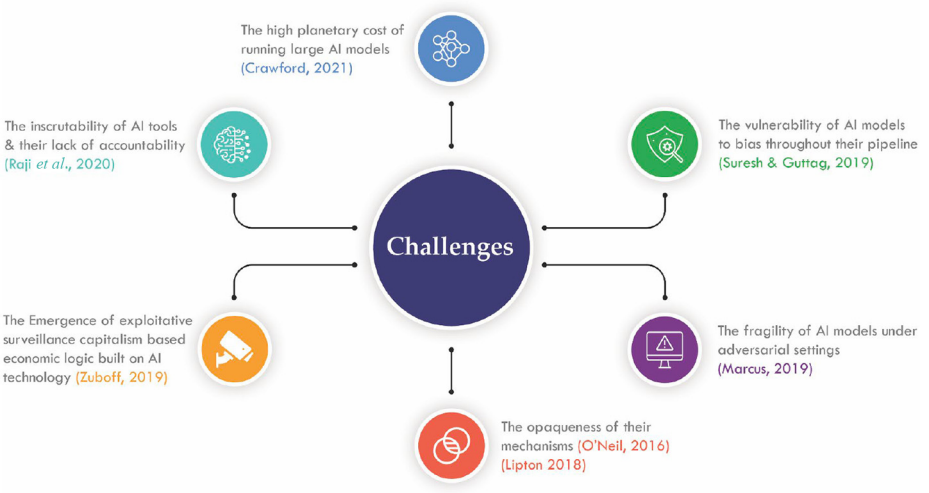

Figure 1.
Challenges and downsides associated with AI

mental health (stress, loneliness), politics and elections (through misinformation and propaganda), as well as systemic oppression (for instance, amplification of racism). Other challenges include "deepfakes," the military use of AI technology for automated warfare, and the disruption to employment by AI systems.

While massive digitization has been adopted in developing economies, there is a risk that the benefits may be undone through concentration of resources and profits in certain quarters preventing the entire society and the whole of humanity to benefit. For instance, the development of AI technology is currently concentrated in only some quarters and driven by certain demographic groups, a situation that AI expert Kate Crawford calls the *AI's white guy problem*. Like all technologies, AI technologies also incorporate and reflect the values of its creators. As Kate Crawford writes in her article [2]:

> [. . .] inclusivity matters – from who designs it to who sits on the company boards and which ethical perspectives are included. Otherwise, we risk constructing machine intelligence that mirrors a narrow and privileged vision of society, with its old, familiar biases and stereotypes.

In response, some authors are calling out for the adoption of "decolonial" thinking (Mohamed *et al.*, 2020) to minimize the hegemony of any one group in directing the trajectory and development of AI. Critical questions need to be asked about where development is happening, and who is doing it, and to investigate power and culture embedded in a system. It is necessary for promoting fairness, justice and beneficence for all that marginalized groups need to be included and given opportunities to influence the process.

AI techniques are heavily dependent on the data provided, expressed concisely by the adage "garbage in, garbage out." The collection of quality data often poses a significant challenge for many developing economies, due to its dependence on resources, manpower, monitoring and processes. Data collection by government comes with its own pitfalls. As the claims of developmental progress often depends on publicly released numbers, such data is often influenced, massaged or doctored, and may only be partially available, and when particularly damning, not released at all. The usefulness of such data in the effective steering of the government's and development sector's policy decisions thus becomes questionable. A startling reminder of the unreliable nature of data in developing countries can be seen in an example described by Latif *et al.*, 2019, who discussed how the GDP of Nigeria and Ghana was recalculated to result in a whopping overnight correction of more than 60%. Experts have shown that bias, far from being a rare occurrence that creeps into the system through the training data, can afflict all parts of the AI pipeline (Suresh and Guttag, 2019).

3. State of practice in artificial intelligence ethics

The growth of AI technology powering numerous touch points in human lives has also exposed us to the risks associated with it. As these have become more apparent, the past few years have seen a rapid increase in development of ethical AI guidelines (Jobin *et al.*, 2019) to ensure AI remains socially beneficial. Consisting of principles, values and best practices, these have been issued by private organizations, public bodies and research institutes, each with slightly different priorities, but with an apparent convergence on five key principles: namely, justice and fairness, transparency, non-maleficence, autonomy and responsibility.

Other investigations into these core principles have shown that while some of the principles are shared with previous bioethics guidelines, the principle of explanability or explicability is especially important for AI. Mention of either accountability, explanability, transparency or interpretability is found in all guidelines involved in the study (Floridi and Cowls, 2019), capturing the convergence on opacity of some forms of AI. Explanability is

also a crucial enabler for the other four principles, as it helps explore why something happening the way it is, what are the possible outcomes, what AI would do if substituting a human. At the same time notable differences exist regarding how these principles are to be interpreted and implemented (Jobin *et al.*, 2019).

While there have been efforts made to define what AI ethics are, the challenges of implementation remain unsolved. The abstract nature of ethics guidelines makes them difficult for developers to adopt in practice. This has led to a disconnect between the AI ethics community and AI practitioners. This has necessitated a shift from a mere description of ethics to an application of ethics. In other words, the emphasis is moved back from code to humans who become the active recipients of AI ethics who experience ethics not as an abstract concept but as a concrete reality.

Another challenge in making AI developers and corporations accountable is that the guidelines are themselves developed by the technology companies. Various stakeholders have express unease with the concept of self-policing due to the likelihood of relaxation of ethical standards when some economic gain is at stake. Currently, for the most part, there is no central enforcement authority within a country or internationally, which can enforce ethical guidelines in any serious way.

In earlier work on publicly available AI tools, Morley *et al.* (2019) suggest a typology that aims to bridge the gap between five core principles and real-world practices. While themes describing beneficence, non-maleficence, autonomy, justice and explicability are prevalent in AI ethical documents, they fail to generate actual changes in the design of algorithms, giving rise to prevalent ethics shirking by businesses. In essence, practice remains divorced from principles. Against this backdrop, the proposed framework invites developers to consider the five ethical issues at each step of the development process. Interdisciplinary expertise here is used to fully be able to translate principles into practices. Although including ethical tools and frameworks may add to overhead for AI businesses, the threats posed by short-termism are too significant to dismiss. While this typology has limitations, it is meant to serve as a point of departure where developers can access relevant tools and methodologies and initiates a rationalization process applying, evaluating and reapplying to ensure ethically aligned results.

There is also a need to train young AI scientists and students on ethical questions surrounding their use. Some educators have adopted "deep tech" approach to tech ethics (Ferreira and Vardi, 2021) integrating standard elements of technology ethics into a more holistic outlook that also embraces sociology, politics, social justice and development of potential socio-technical solutions. Others have worked to incorporate diverse value systems in developing a syllabus for teaching ethics, together with secular ethics frameworks (Qadir and Suleman, 2018; Hughes *et al.*, 2020).

Operationalizing ethical AI has also been at the core of the fight between regulators and US tech giants who seek to monetize user data by serving hyper-personalized ads. The General Data Protection Regulation represents a landmark shift in this regard towards holding tech giants accountable, with not only a unification of laws across the European Union (EU) but also stricter penalties which cannot be ignored. While a key point of departure is the self-governance model in the USA compared with the supremacy of individual privacy in the EU, it cannot be denied that the Snowden revelations and the Cambridge Analytica scandal have increased awareness in the public of the risks of breaches and potential misuse (Houser and Voss, 2018).

The issue of operationalizing AI ethics guidelines has also been tackled by using custom checklists co-created with practitioners (Madaio *et al.*, 2020). This approach ensures that the checklists are grounded in practitioner needs, as they traverse development and deployment

lifecycles. Checklists have not always been successful in other domains, and co-creation is vital to prevent misuse, increase adoption, and hence increase adoption and implementation of guidelines in practice. Other findings were that organizational culture and leadership buy-in is also important for success of checklists.

4. Accountable human-centered artificial intelligence: promising directions

As the field of AI matures and becomes central in the lives of people, it becomes ever more important to be cognizant of issues pertaining to robustness, fairness, interpretability and safety. While in the early stages of development it was acceptable to have a more practical outlook, with rapid progress facilitated by weak controls, we can do so no longer. Ensuring auditability has become necessary to identify risks before they are deployed in production and have already caused harm.

Some studies have advocated rigorous adoption of traditional scientific methodology of experimentation and hypothesis testing in the domain AI research, where it has traditionally been lacking (Forde and Paganini, 2019). Showing how statistical testing in high energy physics has been adopted successfully, they proceed to develop the analogy with AI implementations and research to demonstrate the potential.

Others have worked to design frameworks for making internal audits impactful, thereby increasing accountability for AI applications. Audits are often slow, methodical, and meticulous, and often diametrically at odds with the rapid development approach today but have become necessary in high-stakes domains using AI. A previous work (Raji *et al.*, 2020) proposes a customized internal audit framework which is interdisciplinary and breaks down the process into digestible parts, while requiring essential relevant documentation from all of audit, product and engineering teams.

In the following subsections, we introduce some promising directions for developing human-centered, which are visually depicted in Figure 2.

4.1 Explainable and interpretable artificial intelligence

In recent years, the topic of Explainable AI has seen rapid increase in importance (Arrieta *et al.*, 2020). A key factor to this increase in recognition and visibility is that it is touted as a potential solution to the problems posed by new advances in rapid AI development. In recent times, deep neural networks and ensemble techniques have begun to be known as black-box models, with millions of parameters and hundreds of layers. As these models become more widely adopted in important domains including finance, health care and

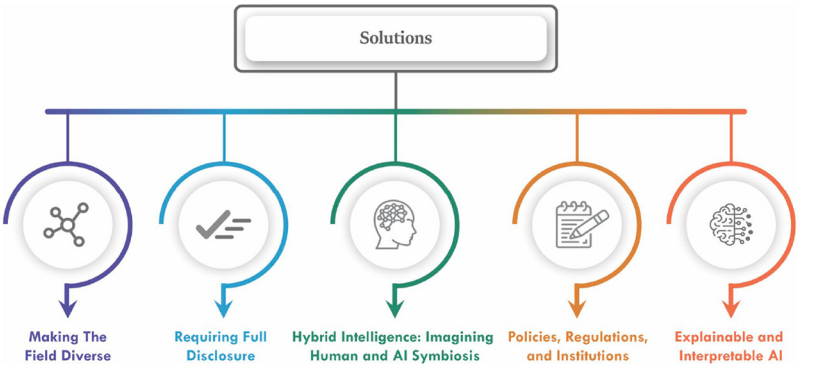

Figure 2.
Promising directions for developing human-centered AI

justice, knowing the decision-making rationale behind AI-powered decisions becomes essential before their implementation. In this regard, apart from explaining black-box models, newer interpretable techniques can be developed. It is also recommended that there be more openness and transparency (Piano, 2020).

Factsheets about AI services which include product lineage and the safety and performance testing it has undergone have also be suggested by some studies to foster trust (Hind *et al.*, 2018). The important task of "interpretability" of AI models is complicated by the fact that it is often referred to either in abstract terms, or in terms which are not in harmony with other definitions in academic work. As the principle of interpretability is key to earning "trust," the various facets of what trust means in this context add further nuance to this discussion. Studies also grapple with ideas of *post-hoc interpretability* – ways humans seek to explain decisions without interpreting the mechanisms which make models work (Lipton, 2018).

Analysis conducted on explanability in deployments (Bhatt *et al.*, 2020) also yields insights on challenges and successes as they are seen in practice and suggest future paths to adopt. Significant limitations of explanability in deployment include lack of causal inference (Marcus and Davis, 2019; Pearl, 2019), the increased latency in calculating and showing explanations, and the risk of misleading correlations reflected in model explanations.

4.2 Policies, regulation, institutions

Policies and regulations are essential components in any strategy to harness nascent technologies and steer them towards desired outcomes. AI has been called "the most important general-purpose technology of our era" by Harvard Business Review [3], and the previous hands-off approach, dealing with problems as they arise, has proven to be a flawed approach. Industry giants, including Google CEO Sundar Pichai [4], have now also begun to call for sector-by-sector regulation of development of AI applications. It is therefore in our best interest to address these challenges and risks proactively rather than retroactively.

Roberts *et al.* (2021) have highlighted automation initiatives can yield great economic benefits using AI. Conversely, this can increase inequalities already present in the society and decrease support for the government. In his book on this subject (Toyama, 2015), Toyama has written about his experiences in India on a decade long ICT4D project. His central argument is on the laws of amplification and how they are typically understood in this space. According to Toyama, technology and gadgets do not "flatten" the playing field but enable those with better skills and knowledge to get higher leverage. He goes on to suggest the needed analog complements comprising a holistic "heart, minds and will," and "intrinsic growth." His insights sheds light on not only the white savior complex but also how the tech component is lacking in essential complements when it comes to practical scenarios.

To this end, policymakers, international organizations, professional bodies are now collaborating to develop a safer AI environment. Academics, as well as public and private sector initiatives, have joined hands in contributing to AI research. Various international states have also launched ambitious strategies to advance development and commercialization of AI while sustaining economic competitiveness. In other words, various actors are taking part in enriching their understanding of AI, potential harms and how it should be regulated.

To devise effective regulations, it is paramount that state structures work with societal stakeholders as well as international actors. As nations begin to devise strategies, their approaches begin to come into conflict with one another. Moreover, their ability and willingness to control the effects of their actions in foreign jurisdiction is limited. In other words, such policies no longer remain a national concern rather they take up a transnational

character. In such cases, it becomes essential to align the aims of various nations when attempting to devise regulations.

Regulation is not just restricted to rules. Rather, it is a system which begins with identifying an anomaly to devising an adequate response coupled with effective supervision, enforcement, continuous assessment and adaption of regulatory regimes. Put simply, it is a system where international, national, and regional workers work towards common goals. It must be a decentralized process of input from multiple stakeholders rather than a purely state driven initiative.

Earlier work by Erdélyi and Goldsmith (2018) suggested the formation of an International Artificial Intelligent Organization (IAIO), which would act as an international forum for discussion and policy development. Through such a forum, policy makers would benefit from the interdisciplinary expertise of range of stakeholders from public sector, industry and academia to develop a system of regulation. The following characteristics would be useful:

- Owing to the novelty of the situation, flexible cooperation arrangements would be better suited rather than binding commitments, as these parties begin to engage with uncertain issues, which may change rapidly.
- As weaponized AI technologies and certain mining practices have high sovereignty costs attached, matters of utmost importance to national security, states will be reluctant to delegate decision making authority to IAIO. Delegation of these may be relegated to later steps in the process.
- While a collective control of information is the goal, states will be reluctant to share their information on leading edge AI technologies. This can be facilitated by using soft law instruments such as guidelines, standards and nonbinding recommendation, which could be turned into concrete legal instruments over time.
- Low initial contracting costs are better suited because of the states would be beginning to get familiar with one another and reaching speedy negotiations and concluding agreements would be easier.
- It should have minimalist administrative functions with the "less is more" approach. The focus must be on maturing the purpose of the organization, its membership, and the issues at hand.
- The organization must invest time and energies in establishing shared interests, ideas, cooperation through routine management rather than opting for a crisis management. Put simply, IAIO, in the initial stages, may begin as informal intergovernmental organizations, display a relatively low level of institutional formality, use soft law and assist national policy makers in developing AI regulations.

In 2017, China released a document outlining China's AI policy objectives named, the "New Generation Artificial Intelligence Development Plan". It delineated policies focused on international competitiveness, economic growth and social governance and making China a key player in the worldwide AI arena by 2030. Similarly, while social issues such as pollution and standard of living are being addressed, people's privacy is being compromised. Individual privacy is being breached in the name of public good where group is given benefit over individual. The government collects personal data wherever and whenever it considers right for policy objectives. The same is true in health care, where data is shared with various government bodies without individual consent. While this may be perceived as social welfare, it does not absolve the state of breach of privacy and poor medical ethics. The document serves to demonstrate that while AI policies are key to China

and its internal needs, it is cognizant of shortcomings and ethical concerns. A more detailed analysis of the Chinese AI approach and a comparison with other approaches is provided by Roberts *et al.* (2021).

4.3 Requiring full disclosure

As the reach and impact of AI systems expands, there are several areas where there is still little to no transparency. Increasing access to these would contribute to increased democratization of responsible AI technology and accountability and show why certain AI works the way it does.

Research on the aspect of trained AI models and their under-performance with certain types of subjects has led to studies (Mitchell *et al.*, 2019) showing the utility of model fact sheets (model cards) with details of the model. This would allow practitioners to compare candidate models for deployment along not only performance metrics but also across ethical, fairness and inclusion scales. The model card would include information on model details, intended use, evaluation and metrics, training and test data, ethical considerations as well as recommendations and limitations. These also consider important at-risk intersections of society.

Others have sought to document datasets by developing standardized processes for developing datasheets (Gebru *et al.*, 2018), which would document the collection process, motivation, composition, and recommended uses pertaining to it. As AI models based on machine learning are trained on data sets, selection of a data set with a similar context and without biases is crucial. Mismatches can be especially harmful if the resulting AI service is deployed in a high-stake environment such as criminal justice, finance, hiring or medicine. The proposal that each data set be accompanied by a datasheet aims to increase transparency, mitigate biases, facilitate reproducibility of results and simplify selection of data sets for diverse uses.

There has also been renewed stress on improving reporting of results by improving reproducibility. Work by Dodge *et al.* (2019) has also explored computation budget as a method of equalizing across different models and showed how accounting for these would have impacted results in recent publications. This brings us back to the importance of transparency and disclosure.

4.4 Hybrid intelligence: imagining human and artificial intelligence symbiosis

Despite the impressive success of AI in various domains, we are nowhere close to Artificial General Intelligence (AGI). The success of AI has been mostly in narrowly defined specific tasks. It is well known in the AI community that the tasks amenable for artificial and human intelligence may be quite divergent. For instance, the *Moravec Paradox* states that it is quite easy for computers to match or overshadow humans in intelligence tests or in logical games such as playing checkers but very difficult for AI to have the dexterity of toddlers when it comes to perception and mobility (Dellerman *et al.*, 2021). The strength of human intelligence is in the intuition and common sense and the strength of machine intelligence is in analysis and computation.

Human beings have a clear advance in general purpose intelligence and the case for developing hybrid intelligence (human intelligence augmented by AI) is strong as argued by Kamar (2016) and Dellerman *et al.* (2021). Keeping humans in the loop is also important since human beings, and not the AI algorithms, are morally responsible. Keeping humans in the loop can avoid the Value Alignment problem that plagues AI systems (Christian, 2021).

Experts in various domains are converging to the realization that hybrid AI systems (with the right function split) are more reliable than pure AI systems. For instance, the

company Locomation is used human-guided autonomy for developing reliable self-driving trucks. The idea is to use a two-truck convey with a lead truck and a follower truck with only one driver being involved at one time in the lead truck whereas the other one rests off the clock. Till full autonomy in reliable cyber physical systems is possible, and AI matures to AGI, it is likely that we will have to use such AI and Human Intelligence symbiosis.

One way of augmenting human intelligence is to rely on crowdsourcing rather than on individuals. This improves the decision-making through the wisdom of the crowds, as the biases of different individuals are neutralized (Surowiecki, 2004). It should be ensured that ethical norms and issues are followed in crowdsourcing activities and that the effectiveness of crowdsourcing is not blunted by lack of diversity or bias.

AI can help improve human intelligence in two main ways: it can be used to automate tasks for instance, the through machines; it can provide decision support to humans who can act inconsistently and sub optimally by violating probability rules. As discussed by Dellerman *et al.* (2021), we can better achieve complex goals by combining human and AI, thereby returning improved performance compared to what each could have managed.

4.5 Making the field diverse

For AI to be human-centered and humanity-centered, it is important that it embraces the diversity of human beings. AI systems are not divorced from their socio-cultural settings. Although they may appear objective and neutral on the surface, if unchecked, AI systems have the potential to reinforce societal biases putting vulnerable groups of individuals at a further disadvantage. It therefore becomes imperative to have a diverse pool of data. As an example, in applications like automated melanoma detection from skin images which detects melanoma, a type of skin cancer, the harm becomes life threatening if the system fails to recognize it in certain skin tones. In other words, AI systems have far-reaching consequences that manifest themselves within the human societies.

To avoid these possible pitfalls, adopting an interdisciplinary approach becomes essential whereby AI can draw from fields which have longer histories of dealing with human sensitivities. Social sciences such as sociology, psychology, economics, history and anthropology have proved to be beneficial for holistically analyzing the complexities of the human subject, and the inherent biases that exist within society and the insights they bring can be enrichening and complementary to purely technical approaches (Dignum, 2020; Sloane and Moss, 2019).

To begin with, AI must focus on methodologies for data collection and annotation. This is so because a fault in data collection basics inevitably translates into issues of data set composition and resultant outcomes of AI. While data collection, especially annotation, has garnered growing interest of researchers, ethical questions such as consent, privacy, inclusivity, power and transparency remain largely unexplored. Earlier work (Jo and Gebru, 2020) explores at length how document collection practices adopted by archives and libraries can positively inform data collection methodology of AI systems and their application in AI settings. These have been codified in five main approaches: consent, inclusivity, power, transparency and ethics/privacy. Steps must be taken at both the macro level (community) and the micro (individual practitioner) to develop professional industry-wide standards for data collection and annotation to realize key goals.

Considering diversity and inclusion is also relevant in subset selection while accounting for differences in social power and access dynamics. Metrics to calculate these have been

highlighted in a key work on this topic (Mitchell *et al.*, 2020). Addressing diversity concerns will help us eliminate, or at least lessen, historical and representational bias. The former signifies structural and empirical inequalities intrinsic to society, for instance the historical lack of women presidents in various countries while the latter encompasses existing barriers within a society which inhibits a group's ability to be digitized and preserved. This automatically skews the data as a section of a society is absent.

By tackling these issues, AI systems can become more inclusive and diverse. Diversity, however, entails having variety in the representation of human subjects with respect to their socio-cultural background. For instance, a diverse pool of data would cover characteristics such as race, age, gender, sexual orientation, etc. within a group of subjects. Inclusivity moves a step further; it entails an individual accessing a diverse data feeling a sense of belonging and are therefore able to benefit from that diversity. In other words, it represents an individual user within an instance. This translates into better alignment between a user and the options available to them in a set. For instance, a diverse data for images of construction workers would display both men and women as workers. However, an inclusive data would mean that both male and women are represented as working in a modern realistic setting rather than one gender being shown as toys, clipart, etc. In the latter scenario, the gender being presented as clipart might not feel connected to the image and would feel disadvantaged for lack of representation in the real world.

5. Conclusions

AI has the potential to benefit the entire society, but there remains an increased risk for vulnerable segments of society who may already be under increased structural challenges, of the harms of many of these systems. Our main contribution in this paper is that we comprehensively review the ethical and social challenges related to the practice of AI and discuss potential solutions, both technical as well as non-technical, which can pave the way for a more pro-social future for AI. The downsides of AI include the vulnerability of AI model to bias and adversarial attacks, their opaque black-box nature, the lack of transparency, the high cost of running large AI models and the emergence of an exploitative economic ecosystem around AI built on surveillance capitalism. In this paper, we have reviewed various promising directions being explored for developing human-beneficial accountable AI including bringing more diversity to the field, requiring full disclosure, exploring hybrid intelligence, strengthening institutions, regulations and policies and emphasizing on the development of explainable and interpretable AI. The challenge of developing accountability for AI, in an age where tech behemoths rule on the back of unfettered AI applications is not simple. However, in this high-stakes battle, it remains paramount to remember that failure to adopt these in time may mean relinquishing aspects of basic human rights and human freedoms to these corporations.

Notes

1. Available at: https://ledger.humanetech.com/
2. Kate Crawford, New York Times, available at: www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html
3. Available at: https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence
4. Available at: www.ft.com/content/3467659a-386d-11ea-ac3c-f68c10993b04

References

Ali, A., Qadir, J., Ur Rasool, R., Sathiaseelan, A., Zwitter, A. and Crowcroft, J. (2016), "Big data for development: applications and techniques", *Big Data Analytics*, Vol. 1 No. 1, pp. 1-24.

Arrieta, A.B., D´Iaz-Rodr´Iguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garc´Ia, S., Gil-L´Opez, S., Molina, D., Benjamins, R. and Chatila, R. (2020), "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI", *Information Fusion*, Vol. 58, pp. 82-115, doi: 10.1016/j.inffus.2019.12.012.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M. and Eckersley, P. (2020), "Explainable machine learning in deployment", *in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 648-657, doi: 10.1145/3351095.3375624.

Buolamwini, J. and Gebru, T. (2018), "Gender shades: Intersectional accuracy disparities in commercial gender classification", in *Conference on fairness, accountability and transparency*, PMLR, pp. 77-91.

Christian, B. (2021), *The Alignment Problem: How Can Machines Learn Human Values*?, Atlantic Books.

Crawford, K. (2021), *The Atlas of AI*, Yale University Press, doi: 10.12987/9780300252392.

Dellerman, D., Calma, A., Lipusch, N., Weber, T., Weigel, S. and Ebel, P. (2021), "The future of human-AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems", arXiv preprint arXiv:2105.03354.

Denning, P.J. and Denning, D.E. (2020), "Dilemmas of artificial intelligence", *Communications of the ACM*, Vol. 63 No. 3, pp. 22-24, doi: 10.1145/3379920.

Dignum, V. (2020), "AI is multidisciplinary", *AI Matters*, Vol. 5 No. 4, pp. 18-21.

Dodge, J., Gururangan, S., Card, D., Schwartz, R. and Smith, N.A. (2019), "Show your work: improved reporting of experimental results", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185-2194, doi: 10.18653/v1/D19-1224.

Erdélyi, O.J. and Goldsmith, J. (2018), December. "Regulating artificial intelligence: Proposal for a global solution", in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 95-101, doi: 10.1145/3278721.3278731.

Ferreira, R. and Vardi, M.Y. (2021), "Deep tech ethics: an approach to teaching social justice in computer science", in *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, pp. 1041-1047, doi: 10.1145/3408877.3432449

Floridi, L. and Cowls, J. (2019), "A unified framework of five principles for AI in society", *Harvard Data Science Review*, Vol. 1 No. 1, doi: 10.1162/99608f92.8cd550d1.

Forde, J.Z. and Paganini, M. (2019), "The scientific method in the science of machine learning".

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daum´e, H., III. and Crawford, K. (2018), "Datasheets for datasets", in *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning, Stockholm, Sweden, PMLR 80*.

Helbing, D., Frey, B.S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., . . . Zwitter, A. (2019), "Will democracy survive big data and artificial intelligence?", in *Towards Digital Enlightenment*, Springer, Cham, pp. 73-98.

Hind, M. Mehta, S. Mojsilovic, A. Nair, R. Ramamurthy, K.N. Olteanu, A. and Varshney, K.R. (2018), "Increasing trust in AI services through supplier's declarations of conformity", Vol. 18, pp. 2813-2869.

Houser, K.A. and Voss, W.G. (2018), "GDPR: the end of google and Facebook or a new paradigm in data privacy", *Richmond Journal of Law and Technology*, Vol. 25, p. 1.

Hughes, J., Plaut, E., Wang, F., von Briesen, E., Brown, C., Cross, G., Kumar, V. and Myers, P. (2020), "Global and local agendas of computing ethics education", in *Proceedings of the*

2020 ACM Conference on Innovation and Technology in Computer Science Education, pp. 239-245.

Jo, E.S. and Gebru, T. (2020), "Lessons from archives: Strategies for collecting sociocultural data in machine learning", in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 306-316.

Jobin, A., Ienca, M. and Vayena, E. (2019), "The global landscape of AI ethics guidelines", *Nature Machine Intelligence*, Vol. 1 No. 9, pp. 389-399, doi: 10.1038/s42256-019-0088-2.

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

Latif, S., Qayyum, A., Usama, M., Qadir, J., Zwitter, A. and Shahzad, M. (2019), "Caveat emptor: the risks of using big data for human development", *IEEE Technology and Society Magazine*, Vol. 38 No. 3, pp. 82-90, doi: 10.1109/MTS.2019.2930273.

Lipton, Z.C. (2018), "The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery", *Queue*, Vol. 16 No. 3, pp. 31-57, doi: 10.1145/3236386.3241340.

Madaio, M.A., Stark, L., Wortman Vaughan, J. and Wallach, H. (2020), "Codesigning checklists to understand organizational challenges and opportunities around fairness in AI", *in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1-14.

Marcus, G. and Davis, E. (2019), *Rebooting AI: Building Artificial Intelligence we Can Trust*, Vintage.

Mitchell, M., Baker, D., Moorosi, N., Denton, E., Hutchinson, B., Hanna, A., Gebru, T. and Morgenstern, J. (2020), "Diversity and inclusion metrics in subset selection", in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 117-123.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T. (2019), "Model cards for model reporting", *in Proceedings of the conference on fairness, accountability, and transparency*, pp. 220-229.

Mohamed, S., Png, M.-T. and Isaac, W. (2020), "Decolonial AI: decolonial theory as sociotechnical foresight in artificial intelligence", *Philosophy and Technology*, Vol. 33 No. 4, pp. 659-684.

Morley, J. Floridi, L. Kinsey, L. and Elhalal, A. (2019), "From what to how. an overview of AI ethics tools, methods and research to translate principles into practices".

Newport, C. (2020), "When technology goes awry", *Communications of the ACM*, Vol. 63 No. 5, pp. 49-52, doi: 10.1145/3391975.

O'Neil, C. (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown.

Pearl, J. (2019), "The seven tools of causal inference, with reflections on machine learning", *Communications of the ACM*, Vol. 62 No. 3, pp. 54-60, doi: 10.1145/3241036.

Piano, S.L. (2020), "Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward", *Humanities and Social Sciences Communications*, Vol. 7 No. 1, pp. 1-7.

Qadir, J. and Suleman, M. (2018), "Teaching ethics, (Islamic) values and technology: Musings on course design and experience", in *2018 7th International Conference on Computer and Communication Engineering (ICCCE)*, IEEE, pp. 486-491, doi: 10.1109/ICCCE.2018.8539286

Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020), "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing", in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33-44, doi: 10.1145/3351095.3372873.

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V. and Floridi, L. (2021), "The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation", *AI and Society*, Vol. 36 No. 1, pp. 59-77, doi: 10.1007/s00146-020-00992-2.

Sloane, M. and Moss, E. (2019), "AI's social sciences deficit", *Nature Machine Intelligence*, Vol. 1 No. 8, pp. 330-331.

Suresh, H. and Guttag, J.V. (2019), "A framework for understanding unintended consequences of machine learning".

Surowiecki, J. (2004), *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business*, Doubleday

Toyama, K. (2015), *Geek Heresy: Rescuing Social Change from the Cult of Technology*, PublicAffairs.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S. and Nerini, F.F. (2020), "The role of artificial intelligence in achieving the sustainable development goals", *Nature Communications*, Vol. 11 No. 1, pp. 1-10.

Zuboff, S. (2019), *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Profile books.

Corresponding author
Junaid Qadir can be contacted at: jqadir@qu.edu.qa