Putting Ethics at the Core of AI: How Do We Build Inclusive and Intelligible Intelligence?

Daniel Kirby and Dr. Suay M. Ozkula
The University of Sheffield

Executive Summary

Artificial Intelligence (AI) forces humanity to question what is human and what is machine, and generates concern over the erosion of such a distinction entirely. Ethical considerations are essential in this debate because human agency is at stake. The ethical guidelines drawn out for AI today can regulate its impact on future affairs, and thereby shape its role in the social world. During the 2019 World Summit on the Information Society (WSIS 2019), a range of stakeholders came together to discuss the ethical dimensions of AI and initiated a dialogue that centred on issues of impact, marginalisation, trust and intelligibility. This policy brief outlines some of the key ethical issues addressed at the summit. Foregrounding the risks that AI presents for marginalising specific social groups, this policy brief will suggest that there is a need for increased transparency when it comes to designing AI, as well as for critical data literacy when assessing AI trustworthiness.

AI For Good: A Dialogue Shift

Artificial Intelligence has been of interest to the International Telecommunications Union (ITU) for a number of years now. What makes WSIS 2019 distinctive, however, is an increased emphasis on the ethics, trust and transparency of AI. This dialogue shift has become clear by the emergence of a separate ITU summit named 'AI for Good'. The framing of this secondary event alone contains within it a moral question: if we want to develop AI for good, what is AI being developing for at present? Moreover, what does it mean to develop AI for bad?

To some extent, the WSIS has already started to address some of these

questions. One significant aspect of the summit's debates has been the question of where human values in AI lie. The discussions highlighted that there has been burgeoning awareness that AI systems are not value-neutral. Humans build these technologies as an enhancement of their organic capabilities. Much like a prosthesis, these technologies are created to fit human needs and demands. AI systems therefore have human bias embedded at their core, and programmed within their code. These systems learn from existing human behaviours, patterns and structures, which are inextricably linked to hierarchies of inequity, and therefore reproduce them.

AI for Whose Good?

As a result, AI systems may exhibit racial and gender biases, as was the case with Amazon's sexist AI: an automated recruitment tool designed to sort through CVs and manage the company's enormous recruitment pool. This AI was trained using data gathered from Amazon's mostly male workforce. It therefore learned to penalise CVs that included the word 'woman' within the text. One of the consequences of this machine-learned behaviour was that graduates from two all-women colleges received lower scores for their applications.

Another example of AI bias is the tendency for facial recognition technologies to disproportionately mis-gender black women within 3 commercial gender classifiers sold in API bundles by Microsoft, IBM and Face++. What this highlights is the need to address concerns that sit at an intersection between algorithmic decision-making and human rights (click 'webcast'; click 'floor' on highlighted session; 11:15). Ethics are core to these discussions around AI because such systems are not self-regulating, and, left unchecked, they may reinforce and cement existing digital divides and inequalities. As Sandra Wachter from the University of Oxford points out, "the world is biased, the historical data is biased, hence it is not surprising that we receive biased results". Indeed, without preventative frameworks in place, social prejudice, like any other form of human behaviour, can be absorbed and re-enforced through machine learning.

In response to these problems, we outline (below) a range of transparency-based issues in AI that future policy-making should take into consideration.

Transparency: the First Step Towards AI for Good

The first step to prevent exclusionary AI begins from a technical design standpoint. Peter-Paul Verbeek, appointed member of UNESCO's COMEST, brought attention to this during WSIS session 195 (36:50) by arguing that AI must "be able to explain how it arrived at its conclusions" (37:50) with particular concern regarding the datasets that are used to train AI.

Design transparency is a key element in this discussion. Architects of AI need to be clear about the methodology they have used, the composition of their datasets, and the *reasons* for developing an AI. More than that, there is a need for liability standards, should harmful or exclusionary AI become commercially available to ensure accountability is achieved. This may be difficult given the blackboxing phenomenon, whereby companies guard and mask algorithmic decision-making. Whilst there are already some attempts to crack open AI blackboxes, more action needs to be taken to address this issue. For example, there is a need for more robust regulatory topography - that ensures the necessary transparency required for the development of fairer AI.

Ensuring Transparency

Some progress has already been made in these areas. Amongst organisations addressing ethical AI design, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems Data Literacy is leading a practice-based approach to ethical AI design, with recommendations being published in the recent draft Ethically Aligned Design - written in collaboration with 2000 global experts who specialise in AI. The recommendations include the cultivation of a "safety mindset" when developing AI, which seeks to pre-empt the unintended behaviour that may be exhibited by any AI system.

However, such guidelines may be difficult to implement; after all, these technologies learn from human behaviour, and the implications can therefore be difficult to anticipate. Yet, forecasting is an essential step, given the ethical implications that are at stake by failing to do so. This brings up a secondary, but equally important, question. If developers of AI are unable to predict the consequences of their creation, should they create them in the first place?

Transparent Dangers

Ethical considerations of AI do not begin and end with transparency. In fact, transparency itself has the potential to be ethically problematic should an AI be introduced to a sensitive area. In the area of cybersecurity, for example, transparent AI presents some serious risks. Not only are the ramifications of AI in warfare unpredictable, transparency about AI decision-making in cybersecurity would mean revealing key vulnerabilities in weapons systems and security tech - making cyberattacks easier to carry out.

Bruce McConnell, the Executive Vice President of EastWest Institute, made this point during session 199 (1:01:30) at WSIS 2019: "Where human lives are at stake or affected, and where privacy is affected... we should be careful about applying it in to those situations in the first place, until we understand it a little better" (1:02:40). Transparency then is not the only consideration here. The ethics of deploying AI into sensitive areas must also be a part of the discussion - especially in the absence of proper regulatory frameworks to deal with unpredicted outcomes.

Transparent Marginalisation and the Need for Critical Data Literacy

The issue of where AI should (and should not) be implemented does not just apply to security. AI can also pose the risk of further marginalising specific social groups. This is the case with the recently developed AI 'gaydar', developed by researchers at Stanford University. The study claims that using the software VGG-Face they were able to develop an AI that could "correctly distinguish between gay and heterosexual men in 81% of cases, and in 71% of cases for women". Once again, a transparent methodology in this case could pose risks - given that it could empower exploitative actors, by providing a face recognition toolkit for identifying and targeting LGBT persons.

Trust therefore constitutes a significant but also fickle issue in AI design and implementation, a debate that featured strongly in at the WSIS 2019 talks, with one of the sessions specifically named "What Would It Take to Trust AI?". What discussions during this session failed to address is that literacy should potentially take greater precedent over trust. The AI 'gaydar' is not 100% accurate, but without critical data literacy, this margin of error

may become irrelevant. All that is required is that a person, parent or government *trust*s that this AI is reliable, when not all AI are trustworthy and citizens need to learn to differentiate between them. They need to be *data-literate*.

Building Literacy Begins with Vocabulary

One way of improving critical data literacy for AI is by changing the language that is used to discuss it. Throughout WSIS 2019, a running theme and obstacle was the lack of a singular or agreed definition of 'Artificial Intelligence'. Semantic conflict within this field develops out of disagreements on what it is that constitutes *intelligence*. Critics argue (1:40:07) that current Information Communication Technologies (ICTs) are not close enough the threshold of what we might call 'intelligent', and that many of the technological advances labelled as AI, such as algorithms, are simply developments in automation. Others argue (1:39:08) that intelligent systems already exist, but are understood by few, and have significant impacts on human affairs.

One potential solution for easing semantic tensions in this area is to redefine AI as 'extended intelligence' (EI) (14:58-16:02). This revised term could have significant impacts on literacy and dialogue; as it implies the involvement of a human agent, and could therefore help foster accountability. It is harder to blame the irresponsible design, implementation or commercialisation of EI on unpredicted machine behaviour, given that the term *extended* Intelligence implies that such systems are a mere augmentation - built by people, for reasons and purposes. Such improved terminology would reflect better the biases embedded in machine learning and therefore also help towards improving data literacy.

Conclusion

As the world transitions to greater reliance on EI systems, there is a need for globally agreed frameworks that stretch beyond the reach of law - and set the ethical ceiling (14:00-14:50) for responsible EI development. Ethics is important in debates concerning extended intelligence because, without regulation, they have the potential to widen existing inequalities. Alongside regulation, there is a need for education. (1:13:45), so that

critical data literacy around the implications of EI can be achieved.

Despite the risks, EI does open up a unique opportunity to expose social bias rather than reinforce it. Human bias will always exist regardless of technological intervention. There is the potential to use extended intelligence as a mechanism to map these prejudices, making inequality visible and marginalisation difficult to ignore, as was the case with Amazon, which eventually dropped its sexist EI recruitment tool. For changes like this to occur, transparency is key. In-built prejudices can only be unlearned if AI systems are open to public scrutiny, not hidden or blackboxed.

Third party bodies can help achieve this accountability, by acting as AI watchdogs, and regulating the impacts of automated decision-making on peoples' lives. However, this is a reactive approach to the issue of AI ethics, not proactive. International bodies such as the IEEE, UNESCO and COMEST and will need to continue pushing for a proactive approach by implementing compliance frameworks, and setting standards around the design, implementation and commercialisation of AI technologies. As Dr. Salma Abbasi, Chairperson and CEO of The eWorldwide Group argued, "[i]n the development of AI, private sector companies must be held accountable" (p.35), which will require a holistic approach involving a range of stakeholders and the development of common standards and agreements that fulfil the UN's commitments to human rights and Sustainable development goals (SDGs).

Daniel Kirby is a student on the MA Digital Media and Society at the University of Sheffield.

Dr. Suay M. Ozkula is a Research Associate and University Teacher at the University of Sheffield.