AI & SOCIETY
https://doi.org/10.1007/s00146-023-01661-w

CURMUDGEON CORNER

Ethical AI does not have to be like finding a black cat in a dark room

Apala Lahiri Chavan[1] · **Eric Schaffer**[2]

Received: 12 October 2021 / Accepted: 4 April 2023

Despite (or perhaps because of) all that is being written, filmed, and discussed about artificial intelligence (AI), it is still unclear how AI should be used. How will AI be intertwined with the future narrative of humanity and our planet? Is AI going to follow the path of the creations of Hephaestus, Talos, and Pandora, whose stories filled the imaginations of people in ancient Greece? At the Center for Advanced Study in the Behavioral Sciences at Stanford, Adrienne Mayor described Pandora, a woman, as a type of AI agent. Mayor, who has been researching the history of AI, said, "Her only mission was to infiltrate the human world and release her jar of miseries." Millennia later, how is the interaction between AI and human beings taking shape?

AI philosophers like Yudkowsky, Goertzel, and Bugaj have posited various scenarios such as a sysop, or "systems operator," where an AI is like a benevolent dictator. This form of AI possesses superhuman powers of compassion to ensure that the world is a place that welcomes all life forms and minds who inhabit it. While this form of AI certainly directs its evolution, it also works as a system operator, an AI Buddha of sorts. They also describe an alternative "AI big brother" scenario which visualizes the creation of AI with superhuman capabilities, but with no ability to direct its evolution. Instead, AI's purpose is to preserve the human status quo. Of course, there is the ever-intensifying clash between the 'AI for good of humanity 'and 'AI will destroy the human race' camps' opposing points of view.

In the midst of the ongoing polarization about AI's place in the world, the discourse on AI and ethics has gained impetus. The question that arises is: whose ethics should be used as guiding principles, given that the question of ethics is influenced by cultural values? Discussions about decolonisation and inclusivity as critical considerations when building AI systems are emerging and need to be amplified. We need to be very aware of the not so glorious tradition of following the mistakes of modernism and orientalism, encoding and engraining the centuries-old biases to the emerging intelligent systems according to Kurt Ozenc, author of 'Hey AI, Keep Culture Beautifully weird!.' He also emphasises the need to create more inclusive frameworks which should form the basis of AI systems, so that we can mitigate the possibility of technology fuelled conflicts arising due to the use of hegemonic cultural templates by AI algorithms. Virginia Eubanks in her seminal work 'Automating Inequality' points out how in the new world, inequality and discrimination can be entrenched. What if the algorithm merely bakes in the existing distortions of race and class, making the gulf between rich and poor, white and black, college-educated and manual worker, even more pronounced?

Diverse writing on AI nationalism, data imperialism, and cyber colonisation is emerging even as the stranglehold of data on our lives increases every day. Karen Hao, in her article 'The Problems AI has Today Goes Back Centuries' in the MIT Technology Review notes, while writing on the topic of Algorithmic discrimination and oppression, how the deep societal structures of racial inequality that are the products of history and politics are being replicated through algorithms "trained on data within a racially unjust society".

The increasing importance of the discourse on making AI ethical is evident from the content analysis of 84 publications by Jobin, Ienca, and Vayena in 'The global landscape of AI ethics guidelines'. This led to the emergence of the following 11 overarching ethical values and principles: transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity.

*This paper presents ongoing doctoral research work by the author.

✉ Apala Lahiri Chavan
apala@humanfactors.com

Eric Schaffer
eric@humanfactors.com

[1] Doctoral Candidate, Srishti Manipal Institute of Art, Design and Technology, 14B126, WeWork, 13th Floor, Tower B 247 Park, Hindustan C. Bus Stop, LBS Marg, Gandhi Nagar, Vikhroli West, Mumbai 400079, India

[2] CEO, Human Factors International, Fairfield, USA

Published online: 04 May 2023

1 Missing pieces in designing ethical AI: from abstract values to actionable concepts

Given the increase in efforts to create guidelines, principles, and artifacts to reinforce AI ethics, why are instances of algorithmic biases still on the rise? Apparently, the challenge lies in the difficulty of envisaging how to apply the agreed-upon abstract values forming the basis of the diverse principles and guidelines. As Reid Blackman, author of the recent HBR article, 'A Practical Guide to Building Ethical AI' indicates, "What exactly does it mean to be for fairness? What are engineers to do when confronted with dozens of definitions and accompanying metrics for fairness in computer science literature? Which metric is the right one in any given case, and who makes that judgment?".

1.1 An example of moving from the abstract to the actionable

Professor Kenneth Forbus, in his article 'Creating AI Systems That Take Culture into Account', has written about researcher Morteza Dehghani's work that provides an inspiring example of moving beyond the rhetoric about ethics in AI. Dehghani has developed a computational model called MoralDM, based on the progress seen in cognitive science, in computational modelling of analogy. MoralDM, as Forbus says, "takes a decision problem, stated in simple English, and works through what to do. It uses analogies with culturally specific stories and prior problems to make a decision. Its reasoning can be inspected, including the values identified and their source." It is important to note that MoralDM's decisions change if the stories available to MoralDm that reflect cultural values (e.g., Iranian versus American) change.

This new computational model opens up the possibility of collecting cultural narratives and making them accessible to AI systems, thereby helping model various aspects of a culture. Enabling AI systems to be guided by cultural values and norms would enhance the probability of outcomes that are aligned with the cultures they operate in.

Dehghani's work inspired the authors of this paper to think of the possibility of applying a human-centered perspective to conceptualise an ethical AI design system that can be available for AI designers and developers to use, in its entirety or parts thereof, as needed.

1.2 Ethical AI design system

A design system is described by Hacq, in 'Everything you need to know about Design Systems', as "the single source of truth which groups all the elements that will allow the teams to design, realize and develop a product. So a Design System is not a deliverable, but a set of deliverables. It will evolve constantly with the product, the tools and the new technologies."

Could a human-centered perspective provide a way to build on the discourse about ethical AI and lead to an 'ethical AI design system?' While reviewing the common framework in which AI systems currently make recommendations, we found that there is no ethical input or control of the AI operation. The AI runs without reference to ethical principles, guidelines, or artifacts. The wider ethical intent of the system stakeholders and designers is not integrated into AI operations.

In most AI applications, the AI engine is tasked with making decisions supported by big data and driving a cycle of decision-making and monitoring of results. This allows machine learning to optimize criteria. For example, a bank may want to minimize credit card defaults; thus, they utilize AI technology to optimize credit decisions. AI uses big data and the ongoing flow of credit default results to craft a predictive model.

What if we could have ethical limits embedded in an AI application? An ethical AI design system would at one level be a repository of values, guidelines, and best practices. However, it would also provide actionable components (that can be directly used in the design of AI systems) derived from the various values, guidelines, and best practices. This design system would provide guidance about cultural nuances of the recommended values framework, since values and ethics are influenced by cultural norms.

1.3 The concept of filters: adding equalize filters

Consider a process where the use of big data for machine learning is overlaid with filtering mechanisms such as the 'Equalize filter' which helps operationalize relevant values from a superset of universal values (e.g., the 11 values mentioned earlier). This use of a value-based filtering mechanism avoids biases in the criteria used by the algorithm for recommendations.

1.4 How would the equalize filter work?

Let us consider the scenario of an AI system that recommends whether an applicant for a loan should be approved. Without an equalize filter, the AI algorithm could have biased ways of learning from the available data. The AI system may identify that people of a certain ethnicity have a higher credit risk than people of another ethnicity. If the AI system made this identification very obvious, we would immediately see that it was unethical and possibly illegal. However, the AI system might also find a surrogate variable

(i.e., more difficult to detect because it is not so overt), which makes the same ethnic discrimination, but, in this instance, seemingly based on some objectively observed behavior. For example, it might use the brand of shampoo purchased by people to arrive at a recommendation that in reality would be based on unethical racial discrimination. In the concept of the equalize filter, we mitigate the above-mentioned biases by defining the types of discrimination that are prohibited. For example, ethnicity and wealth can be eliminated as predictors, in all forms. The AI is then constrained to find the predictors of creditworthiness that do not reflect these circumscribed factors.

By eliminating racial and economic factors, we might make credit decisions that reflect a consumer's individual behavior, without any attempt at correlation with their social category, and thus break the negative cycles of discrimination. This is true for determining the credit score and most other AI-based decision-making. Ongoing research on cognitive biases and associated debiasing techniques—drawing from cognitive psychology and applying the results to machine learning—provide further strategies for designing and developing unbiased AI systems.

An ethical AI design system such as this would help mitigate the challenge of envisaging how to apply the agreed-upon abstract values forming the basis of the diverse principles and guidelines and provide ready to plug in components that manifest respect of universal human values (and perhaps going beyond the Anthropocene, even incorporate planetary values), cultural nuances, and individual preferences.

These are hazardous times for our civilization. A wise application of technology is the need of the hour. Thus, we need to intentionally apply AI. AI does not exist in isolation. We need to combine it with additional modalities, such that it is intentionally focused on urgent human needs. It is perhaps the need of the hour to heed Heidegger's advice about understanding the ways of thinking that lie behind technology, so that humans can enter into a "free relationship" with technology by "bringing forth" the evolution of the relationship rather than "challenging forth". Actionable and transparent ways to make ethical AI work for all of humanity and the planet will hopefully make these lines from Pablo Neruda's 'The Watersong Ends' NOT true for how we let AI affect the lives of everyone.

> *'Man turned to his mechanisms and made hideous*
> *His works of art, his lead paintings, his wistful statues of wire,*
> *……And while they arrived on the moon and dropped tools of gold there,*
> *We never knew, children of the slow half—light,*
> *If what was discovered was a new planet or a new form of death'.*

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for superhuman intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? -Editor.

Availability of data and material Not applicable.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.