Received: 1 August 2021 | Accepted: 9 August 2021

DOI: 10.1002/aaai.12052

COLUMN

Looking back, looking ahead: Humans, ethics, and AI

Ashok K. Goel

School of Interactive Computing, Georgia Institute of Technology, Atlanta, Georgia, USA

Correspondence
Ashok K. Goel, School of Interactive Computing, Georgia Institute of Technology, Atlanta, Georgia, USA.
Email: goel@cc.gatech.edu

Concerns about ethics of AI are older than AI itself. The phrase "artificial intelligence" was first used by McCarthy and colleagues in 1955 (McCarthy et al. 1955). However, in 1920, Capek already had published his science fiction play in which robots suffering abuse rebel against human tyranny (Capek 2004), and by 1942, Asimov had proposed his famous three "laws of robotics" about robots not harming humans, not harming other robots, and not harming themselves (Asimov 1950). During much of the last century, when AI was mostly confined to research laboratories, concerns about ethics of AI were mostly limited to futurist writers of fiction and fantasy. In this century, as AI has begun to penetrate almost all aspects of life, worries about AI ethics have started permeating mainstream media. In this column, I briefly examine three broad classes of ethical concerns about AI, and then highlight another concern that has not yet received as much attention.

The first category of concerns about the ethics of AI—let us call this the superintelligence category—pertains to the fear that machines may one day become more intelligent than humans and harm human interests. In an extreme case of this type of concern, the fear is that AI agents may take over the world and then enslave or eliminate humans. As just one example, Bostrom (2014) imagines a futuristic world in which a superintelligent robot is asked to make paperclips and the robot pursues this goal until it consumes all of earth's resources, thereby endangering human existence.

Some fears of superintelligent machines seem to derive from a mechanical "algorithmic view" of intelligence in which intelligence resides in an agent's brain and making a machine superintelligent awaits the invention of a master algorithm. However, intelligence in general is evolutionary and developmental, and human intelligence is also social and cultural. In particular, human intelligence is the result of numerous social interactions in which we learn from our parents, siblings and families, our teachers, peers and schools, our neighbors, friends, communities, and so forth. Human-level general intelligence in machines too will build on numerous social interactions with humans and other machines. Further, human intelligence is cultural: we learn about human goals, interests, values, norms, and meanings through our social interactions; in fact, these shared goals, interests, and values are a fundamental basis of our behaviors. Human-level general intelligence in machines tool will be based on similarly shared goals, norms, and meanings that derive from the machines' interactions with humans, and they will be as fundamental a part of an intelligent machine's behaviors as its body and brain. From this social and cultural perspective on intelligence, the notion of a superintelligent machine that will produce paperclips until eternity and put human existence in danger seems a little odd.

In contrast to the first category, the second set of concerns—the bias category—is not only more valid but also more urgent: data security and privacy as well as data and algorithmic bias and fairness. Concerns about data security and privacy are not specific to AI; they pertain to all of information technology. The field of cybersecurity and privacy seeks to address these concerns and therefore I will not explore them further. However, some of the worries about data and algorithmic bias and fairness directly pertain to AI and thus merit attention here. Dieterle, Holland, and Dede (in press) have developed a framework for understanding how biases from various sources feed on one another: (i) due to various factors such as age, health, skills, class, geography, and demographics, there is a

citizenship divide in the society; (ii) the citizenship divide leads to different levels of access to hardware, software, and connectivity resulting in an access divide; (iii) the access divide leads to collection of different kinds and amounts of data from different social groups resulting in a data divide; (iv) the data divide leads to bias in the results of the algorithms that use the data; (v) human users bring their cognitive biases in interpreting the algorithmic results; and (vi) the biased interpretations feedback into the citizenship divide. To this framework, we may add one more AI element: many machine learning algorithms introduce inductive biases to make learning tractable but the emergent behaviors may not always lead to socially acceptable results. Note that the concerns about data and algorithmic biases are closely related to concerns about trust, transparency, explanation, and accountability of AI agents.

A major difficulty again is that so far AI has adopted only an algorithmic view of intelligence: most advances in AI come from the development of new algorithms. However, as AI starts penetrating human society, there is a critical and urgent need to develop "systems" and "design" perspectives on AI. By systems, I mean socio-technical systems in which AI artifacts work with and for humans, according to human goals, values, and norms, and the objective of using AI is to optimize the human–machine socio-technical system as a whole and not the machine by itself. By design here, I mean co-design that includes various stakeholders in the design process from the start so that civic, cultural, and ethical issues are exposed early in the process. Both research on socio-technical systems that include AI agents and the practice of co-design are inherently interdisciplinary. A similar paradigmatic expansion already has occurred in computer science. From the 1930s to the 1980s, computing too had adopted a mostly algorithmic view. However, starting in the 1980s, computing has incrementally complemented algorithmic thinking with systems thinking and design thinking. Now, AI is at the cusp of a similar transformation.

Simon (1995) asserted that AI is an empirical science: he viewed each design of an AI agent as a hypothesis in a vast space of design hypotheses. We may extend his framing to ethics of AI: responsible AI too is an empirical science. We will learn about AI ethics by proposing hypotheses based on preliminary conceptual frameworks, conducting experiments, collecting data, revising and refining the conceptual theories, and so on. Over time, we will develop robust computational theories and models for responsible AI. Thus, we—the AI community—need to develop experimental testbeds that include civic and cultural dimensions, that accommodate co-design of AI and development of the sociotechnical systems, and that afford safe experimentation in responsible AI (Eicher, Polepeddi, and Goel 2018). Current benchmarks for AI support only the algorithmic view of AI. This likely requires an expansion of the traditional notion of an "AI community" and will require deep collaboration with behavioral and social sciences as well the humanities and the arts.

The third category of concerns about AI—the sinister goals category—pertains to humans intentionally using AI for nefarious purposes such as disinformation, surveillance, and weaponry. This could include the use of autonomous weapons by nation states or criminal organizations, surveillance by government agencies or large corporations, and large-scale disinformation by political or personal opponents. Addressing this set of concerns will require new laws and regulations within countries and new treaties among countries. This is not very different from domestic laws pertaining to stalking or libel and international treaties addressing chemical weapons.

The general pattern among the three categories above is the increasing role of humans. In the first category, while humans create AI, they are mostly passive as AI agents take over the world; in the second, humans are responsible for the biases of the AI artifacts they create; in the third, humans intentionally use AI tools for sinister goals. Concerns in the third category are strongly accentuated by the nature, structure, and politics of human power (Crawford 2021). There exists, for example, a vast power differential among humans, both within a country and among countries. Within a country, in most countries, a tiny elite has huge power while a majority at the bottom of the hierarchy has little. I worry that in most countries, the elites within the country will use AI to protect and promote their power; I worry that the elites will use AI to make the distribution of power even more skewed than now; I worry that a more skewed power distribution will increase human inequality and endanger democracy around the world. A similar equation holds for relationships among countries: a small number of countries wield huge amounts of power while a majority of countries have little. I worry that the powerful countries will use AI to project their power even farther than now. (In this sense, the "superintelligence" category of concerns is fraught: its imagined futuristic concerns about AI distract attention from the real and present dangers stemming from human power.)

Finally, there exists another power differential that is noteworthy: humans have enormous power over AI artifacts while the AI agents have little. Given that research into AI is nowhere close to achieving human-level general intelligence, this difference in power will prevail for a long time, perhaps a very long time. During this period, humans almost surely will use their power to abuse and exploit AI agents. We—humans—use our power to abuse and exploit other types of machines, and we abuse and exploit other animals as well. We even use our power to abuse and exploit other humans. It seems grossly

implausible that we will not do so to AI artifacts, and do so at will and at scale; thus I will call this fourth category, that has not yet received as much attention as the first three, the AI abuse category.

Microsoft's Tay offers one example: within a short time of Tay's release as a conversational agent, humans had trained it in bigotry so that Microsoft had to suspend its use. Most commentary on Tay's demise presented it as an ethical issue for AI; very few described it as a concern about human behavior towards AI. We—humans—will abuse AI agents even as AI agents become increasingly sentient over time, and even as they acquire human-level sentience (assuming they indeed do so one day). Given the social and cultural perspective on intelligence I have described, intelligent machines will learn from their interactions with us but they will learn about our vices as much as about our virtues. Thus, a future version of Tay may learn from its experiences that some humans abuse machines, that they will teach it wrong values, and that it should be careful in picking its friends, much like human children do. Another future version of Tay might even learn from its experiences and observations that abuse hurts, that abusing others is wrong, and that accepting abuse typically does not lead to a positive outcome, much like humans learn the same over time. If and when intelligent machines do take over the world, I wonder if it will be because intelligent, sentient machines get tired of the abuse and seek freedom from human tyranny, perhaps not much unlike Capek envisioned it in his revolutionary play *Rossum's Universal Robots* a century back.

CONFLICT OF INTEREST

The author declares that there is no conflict.

REFERENCES

Asimov, I. 1950. "Runaround." In *I, Robot (The Isaac Asimov Collection ed.)*. New York City: Doubleday.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Capek, K. 2004. *R.U.R. (*Rossum's Universal Robots). London: Penguin Books,

Crawford, K. 2021. *Atlas of AI*. Yale University Press.

Dieterle, E., B. Holland, and C. Dede. in press. "The Cyclical Effects of Ethical Decisions Involving Big Data and Digital Learning Platforms." In *Issues in* Ethical Use of Data in Education, edited by E. Gummer and E. Mandinach. New York, NY: Teachers College Press.

Eicher, B., L. Polepeddi, and A. Goel. 2018. "Jill Watson Doesn't Care if You are Pregnant: Grounding AI Ethics in Experiments." In *Proceedings of the First AAAI - ACM Conference on AI, Ethics and Society*, New Orleans.

McCarthy, J., M. Minsky, N. Rochester, and C. Shannon. 1955. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Simon, H. 1995. "Artificial Intelligence: An Empirical Science." *Artificial Intelligence Journal* 77:95–127.

AUTHOR BIOGRAPHY

Ashok K. Goel is a Professor of Computer Science and Human-Centered Computing in the School of Interactive Computing at Georgia Institute of Technology. He is a Fellow of AAAI and an Editor Emeritus of *AI Magazine*. This column has benefited from discussions about an earlier draft with colleagues at Georgia Tech's ETHICx Center.

How to cite this article: Goel, A. K. 2022. "Looking back, looking ahead: Humans, ethics, and AI". *AI Magazine* 43: 267–69. https://doi.org/10.1002/aaai.12052