

AI Ethics: A Long History and a Recent Burst of Attention

Jason Borenstein, Georgia Institute of Technology

Frances S. Grodzinsky, Sacred Heart University

Ayanna Howard, Georgia Institute of Technology

Keith W. Miller, University of Missouri

Marty J. Wolf, Bemidji State University

Artificial intelligence (AI) ethics has become a hot topic in the popular press and in scholarly writing. In this column, five noted scholars give their opinions on what AI issues will become important in the foreseeable future.

During World War II, a Massachusetts Institute of Technology professor named Norbert Wiener worked on the automatic control of a cannon. In 1948, Wiener¹ coined the term *cybernetics* and wrote about computers:

Digital Object Identifier 10.1109/MC.2020.3034950
Date of current version: 14 January 2021

... we are already in a position to construct artificial machines of almost any degree of elaborateness of performance. Long before Nagasaki and the public awareness of the atomic bomb, it had occurred to me that we were here in the presence of another social potentiality of unheard-of importance for good and for evil.

Terry Bynum² cites Wiener's work in 1948 and in his later book in 1950³ as the start of computer ethics as a scholarly field. Even in this early work, we see the importance of artificial intelligence (AI) issues inside computer ethics, although the formal study of AI is often traced back later to 1955.^{4,5}

Lately, the idea of exploring ethical issues in AI seems commonplace, but it was not always so. We searched Google Scholar for articles or books with a title that includes ("ethics" or "ethical") and ("AI" or "artificial intelligence"). We got the counts shown in Table 1 and Figure 1 for the years 1985 (when



the first such article arrived⁶) through 2020. The count for 2020 is only a partial count, up to when this article was written.

Even though the scholarly literature on AI ethics was limited until the last few years, popular culture was far more engaged in issues related to what we now call AI. The term *robot* is often traced back to a 1920 play by Karel Capek⁷ called *R.U.R.* about automated beings revolting against the human race. Isaac Asimov’s Three Laws of Robotics,⁸ later expanded to four laws, have generated debate for decades. Even a short list of films involving AI⁹ is impressive for their treatment of human interactions with AI: *Metropolis*, released in 1927; *The Day the Earth Stood Still*, 1951; *2001: A Space Odyssey*, 1968; *Westworld*, 1973; *Star Wars*, 1977; *War Games*, 1983; *The Terminator*, 1984; *Short Circuit*, 1986; *Star Trek Generations*, 1994; *The Matrix*, 1999; *AI: Artificial Intelligence*, 2001; *I, Robot*, 2004; *WALL-E*, 2008; *Robot and Frank*, 2012; *Ex Machina*, 2015; *Blade Runner 2049*, 2017; and many others. Similarly, television and steaming movies have taken up these themes with a vengeance.¹⁰ In some sense, scholarly interest is merely catching up to popular culture in its focus on ethical issues and AI.

LOOKING FORWARD: WHAT ARE SOME IMPORTANT ISSUES TO EXPLORE?

When there is a sudden burst of interest (and publications) in a field, it is important to pay attention to the most significant issues and trends. Otherwise, we can be overwhelmed by minutia and spurious, overhyped speculations.¹¹

Many issues in AI ethics are by no means closed questions that have already been authoritatively resolved. Instead, there is a wide array of contentious arguments which we expect to continue for decades. To give *Computer* readers a closer look at three of these contentious issues, we describe three sets of questions that we think are, and

TABLE 1. The counts of Google Scholar citations with (“AI” or “artificial intelligence”) and (“ethics” or “ethical”) in the title.

Year	Count	Year	Count	Year	Count	Year	Count
1985	1	1994	0	2003	4	2012	7
1986	1	1995	0	2004	3	2013	5
1987	0	1996	0	2005	6	2014	12
1988	0	1997	0	2006	2	2015	10
1989	0	1998	3	2007	2	2016	21
1990	0	1999	1	2008	6	2017	45
1991	1	2000	4	2009	2	2018	128
1992	0	2001	0	2010	2	2019	334
1993	0	2002	1	2011	8	2020	342

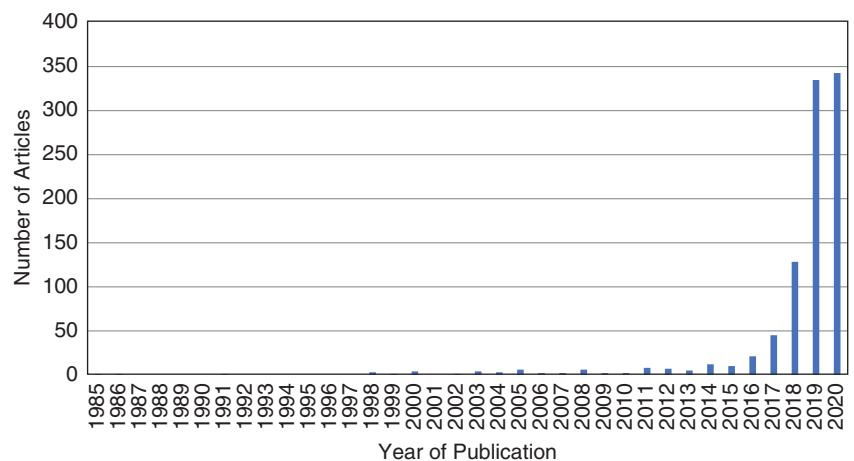


FIGURE 1. The counts of Google Scholar citations with (“AI” or “artificial intelligence”) and (“ethics” or “ethical”) in the title.

will continue to be, significant AI ethics issues in the foreseeable future.

Issue 1. The ethics of exclusion: Deciding who should have a seat at the table when AI systems are being designed (Jason Borenstein and Ayanna Howard)

Many serious ethical challenges are emerging in relation to AI systems. Among them is how to identify and

mitigate different types of biases embedded in the technology.¹² There are also ethical concerns about the disproportionately harmful impacts AI systems are having on the poor, individuals who reflect gender diversity, and people of color.¹³ Moreover, the “dual use” potential of AI systems is a real worry in the sense that a presumably beneficial AI could be maliciously twisted to cause deliberate harm to the public.¹⁴

Along these lines, an AI system that can recommend medical services based on positively maximizing an individual's health outcomes can also be used to turn away those same individuals from receiving comparable benefits to maximize a hospital's revenue stream. Yet in this piece, we want to draw attention to a different issue: one that underlies many of these and other concerns in the realm of AI ethics. It is the overarching question of who should be involved in the process of designing AI systems. Given the widespread impacts that AI systems are having on our lives (both personally and professionally) and how powerful they are in terms of shaping society, drawing attention to who has a

the design process. For example, a designer might train and validate an algorithm's output for results biased against one group, but that doesn't mean the designer has tested for bias against another type of group or even thinks that it is necessary to do so. This is also true when a designer does not consider the intersectionality of identity attributes. A value-free tool in terms of gender may seem fairly accurate with respect to women, but that does not necessarily mean it works well with Black or Asian women.

Like the rest of us, a designer is a fallible, biased human being. Thus, a strategy is needed for sincerely identifying and acknowledging personal

voices, backgrounds, and perspectives that, together, we can make a concerted, collective effort to improve.

So, given this state of affairs, the question thus becomes: "Who does and who should have a seat at the table when AI systems are being designed?" We unpack this question into its two component parts: 1) who is currently involved in designing AI systems and 2) who should be involved. In terms of who is currently involved, it is largely computer scientists and engineers. A typical profile of an AI designer is a person situated in a corporate setting in a relatively wealthy region of the world. Present-day AI designers are predominately White or Asian males.¹⁵

This leads to the harder question, "Who should be involved?" Recognizing the ways in which AI is profoundly changing the world (often in troubling ways), we make the case that the AI design process should be more diverse and inclusive in several senses. This includes striving for more disciplinary diversity among the people taking part in the design process. Sociology, economics, philosophy, the law, gender and race studies, and public policy (among other fields) have valuable insights to share. Diversity in terms of the gender and race of designers is also crucially important; the examples resulting from AI design failures (in part due to the lack of diversity in this sense) are distressing and unfortunately too common, ranging, for instance, from facial recognition failures¹⁶ to health-care treatment errors.¹⁷ Aiming for regional diversity is essential as well. Too often, for example, the Global South finds itself excluded from decisions about emerging technologies generally, and more recently, AI.¹⁸ Finally, it is not just about the designers themselves; it is about who they are interacting with during the design process and when. Potential users and, more broadly, the public must authentically be a part of the picture. A seismic philosophical shift should occur from "What can we design for you?" to "What can we

Those involved in designing computing devices
are typically trained in computer science,
engineering, or a related discipline.

seat at the table during the design process is crucial. We contend that it is one of the most pressing ethical issues pertaining to AI of our time. Deployment and use decisions are, of course, crucial to examine as well, but for our purposes here, we limit the scope to design.

Those involved in designing computing devices are typically trained in computer science, engineering, or a related discipline. Of course, the expertise those disciplines provide is necessary, but not sufficient, to the task of designing AI systems. These specialized realms do not consistently expose students to societal implications interconnected with their future professional work. Moreover, computer scientists, engineers, and others may carry with them assumptions about the value neutrality of technology. The commonly held idea that technology is "value-free" can contribute, as a by-product, to the notion that AI will be better than human decision makers. Yet what such beliefs can obscure is that values (from the designer) are being embedded in technology during

values and overcoming biases (or other shortcomings) when conducting research or designing new technologies. AI cannot achieve the lofty goal of making "better" decisions than humans, assuming that it is even possible, unless diverse voices come together to contribute to the designed solution. Identifying one's own personal biases is difficult to achieve when everyone in the room has a similar background. And it is hard to unwrap one's biases if everyone has similar experiences and similar blind spots. For one, groupthink is likely to result. We (the authors) are not necessarily claiming that designers have bad intent; they probably do not. Many are seeking to promote "responsible computer science," computing for good, or other worthwhile initiatives. But AI and its applications are so complex and reaching into so many facets of our lives that no singular person, discipline, or field is equipped to understand and represent the perspectives that should be included in the design process. It is only when we are confronted with different

design with you?” so that AI is more authentically aligned with what is good for humanity.

And now comes another, even more difficult, question: “How do we effectively manage the process so that those who should be involved have a seat at the table?” Companies and organizations have tried to move the needle over the last few years, but the needle has only twitched a little. Sending the message that “we welcome you” does not resolve the issue if participation is mere tokenism and has no teeth to produce change. There are no simple solutions to be had for this problem. Yet what may help is a cultural shift in thinking, which acknowledges that finding solutions to what originally seemed like a “technical” design challenge requires engaging with individuals and communities who are not traditionally represented.

Issue 2. The ethics of research and development: The training and deployment of AI systems (Marty

J. Wolf and Frances S. Grodzinsky)

Tay was an AI chatbot developed by Microsoft with a goal of learning human speech patterns. Within 24 h, Microsoft researchers had to shut it down when malicious users “taught” Tay to produce and publish anti-Semitic hate speech. In Wolf et al.,¹⁹ we argue that AI, or any software that learns, creates additional risk and places the burden of additional responsibility on not only the software developers who write the AI but also on those who oversee the training of the AI as well. Much of this stems from our concern that it is human subjects who interact with AI systems. This was the case with Tay. Designers did not sufficiently assess the risks to people who came from unleashing it on the open Internet instead of in a closed environment, where it could be closely monitored.

In the United States and many other countries, research on human subjects has a storied past. The notorious 1936 Tuskegee Syphilis Study serves as a marker of how not to conduct research

involving human subjects. It also serves as a reminder that not all experiments should be conducted.

Since that time, a rich set of standards has been developed for research involving human subjects. Policies and procedures exist in those institutions where they can be enforced. U.S. universities that get any sort of funding from the U.S. federal government are obligated to ensure that all research that takes place at the university and involves human subjects meets those standards. In response, most universities around the world have established an institutional review board (IRB) or a similar board that must be consulted

period of time. The case of Tay raises two problems for us: First, what happens when the project comes out of private industry? Is there and, if not, should there be the same kind of oversight? And, secondly, what happens, as in the case of Tay, when the human subjects involved are not specifically defined, but rather general Internet users?

There are two other considerations worthy of note in the context of the development of AI on university campuses. Traditionally, computer science faculty have not engaged in research involving human subjects. Thus, within computer science departments there is

The job of the IRB is to ensure that the research procedures are designed in such a way that subjects are not exposed to any risk beyond that encountered in normal daily life.

at the beginning of a project involving human subjects. Prior to beginning any data collection, the research plan for the project must meet those minimum standards.

The job of the IRB is to ensure that the research procedures are designed in such a way that subjects are not exposed to any risk beyond that encountered in normal daily life. The IRB is responsible for considering physical, psychological, and social risks. Since the IRB must give its approval for the project to go forward, it is in the best interest of those proposing the experiment to consider and address these risks. Privacy and confidentiality are two additional pertinent considerations of the IRB. The protocols and researchers themselves are responsible for ensuring that confidential information, such as names and salient identifying data, are not disclosed outside of the research team. There must also be provisions that protect the confidentiality of subjects whose information is to be retained over an extended

no culture of considering the impact of “technical” computer science research on people, as would be common in social science or biology research, for example. The second problem arises, at least potentially, when a computer science project is presented to an IRB. Those on the panel may likely have insufficient experience with or understanding about the complexities of AI, which is necessary to evaluate the proposal with respect to the risks the subjects and society would be subjected to by the research.

Thankfully, many universities have begun to address these and other shortcomings (for example, understanding confidentiality risks arising from “anonymous” data being combined with publicly available data) in the IRB approval process. It is clearly a work in progress. Yet it serves as a model to address an even bigger problem in the development of AI.

Many of those now working in industry who have come out of computer science programs are still subject to

this technical view of computing. Their view lines up with the definition of weak AI (as opposed to strong AI). In particular, they understand developing AI as functional, that is, performing a task. In that light, they set a goal and write software using standard, well-understood software development methodologies. While these methodologies may be appropriate for standard software, it is not clear that they work well or are considered to be best practice when the software being developed is designed as self-learning, either partially or fully

serious consideration in the ethical analysis. There is a shared ethical responsibility among those who develop the AI, those who train it, and those who deploy it as they position it for use in a global, sociotechnical environment.

At the beginning of this section, we were reminded that the Tuskegee Syphilis Study was a study that should not have been conducted. Given the experimental nature of AI research, development, training, and deployment, it is safe to suppose that not all AI ought to be developed. Each project requires that those involved give serious consid-

decades of overhyped speculation, AI is now presenting us with an amazing array of functioning machines and tantalizing suggestions of amazing advances in the near future. Furthermore, unlike the older fictional stories, we can purchase many of these creations right now, and we are promised increasingly human-like devices in the foreseeable future.

Goaded by recent developments in AI, I think society as a whole and scholars from many disciplines are starting to engage seriously in an increasingly important question, which I will label “Question 0.”

Building with silicon, electricity, and clever engineering, AI has already delivered artifacts that can converse with us, learn with us, and walk with us.

modifiable.²⁰ Microsoft developers who deployed Tay and, subsequently, removed it from the Web did not consider the risks to those merely viewing the posts when Tay unleashed its hateful speech. Should we ascribe responsibility to Microsoft developers for this incident? Tay was quite specific as to its mission to change its behavior in real time according to what it learned from user responses and other information on Twitter and to publicly display its Twitter responses. Risk assessment did not address the larger sociotechnical context of Twitter and its users; nor did it simulate in a closed environment what would happen if learning produced offensive speech. There were not adequate controls for the downstream users as, apparently, Tay learned from all users after deployment.¹⁹ This analysis points strongly in favor of looking at a broader context when designing AI applications. At this point, software developers may not even realize that they need to consider the impact the AI may have on those whose data are used to train it, on those who are subject to its output, or on society as a whole. As previously mentioned by Bornstein and Howard, the “dual use” of AI needs

eration to the question of whether the project even ought to be. Any project must not move forward until there is a clear and convincing argument that humanity will be better off with the AI.

Issue 3. In the future, should we be willing to consider some AI artifacts to be persons? (Keith Miller)

In Greek mythology, Pygmalion sculpts an ivory statue depicting a woman. He falls in love with the statue, and Aphrodite grants his prayer to bring the statue to life. The Greeks did not corner the market on the idea of animating nonliving matter. Jewish folklore includes the golem, and Mary Shelley wrote of Frankenstein’s experiment with reanimation. The subtitle of Shelley’s book is “The Modern Prometheus.” That brings us full circle to the Greeks since Prometheus is the Titan god of fire credited with creating the human race from clay.

Modern AI promises us a transformation similar to these stories. Building with silicon, electricity, and clever engineering, AI has already delivered artifacts that can converse with us, learn with us, and walk with us. After

- › Question 0: Should we consider some AI artifacts, either now or in the future, as persons?

I do not think questions about this issue will be settled quickly or easily; I do think it is vital that we focus on this issue with urgency. I want to distinguish this question from two related, but importantly different, questions:

- › Alternative Question 1: Can a future AI become sufficiently person-like in its behavior and appearance so that we will not be able to easily distinguish between it and a human being?
- › Alternative Question 2: Will society consider some AI artifacts, either now or in the future, as persons?

Both of these alternative questions are interesting, timely, and already being discussed in the literature. Importantly, both of these alternative questions might be answered empirically; if machines exist that routinely pass as human beings (that is, they often pass a physical Turing test), then the answer to Alternative Question 1 is, in my opinion, “yes.” If when we look about and no such machines exist, then the cautious answer to Alternative Question 1 is “at least not yet.”

Similarly, if we look about and observe that either many or most of our fellow humans treat sufficiently sophisticated AI artifacts as if they were persons, then the answer to Alternative

Question 2 is “yes.” If we do not see that happening, then the cautious answer to Alternative Question 2 is “at least not yet.”


Both the alternative questions are questions of observation and description. I think the more important Question 0 is normative. Should we consider some AI artifacts, either now or in the future, as persons? At the heart of this question is a recognition that the designation of personhood is a societal choice, not a scientific classification. That societal choice could be expressed as law, custom, or regulations. Without some societal decision, we cannot devise a definitive physical or behavioral test for personhood: we who are already considered persons have to come to an agreement (probably not a universal consensus) on what other entities, if any, we will allow into our “personhood club.”

Except in the case of cloning humans (for example, the replicants in *Blade Runner*), our “candidates” for personhood have no claim to be of our species. Humans are carbon based, and at least all common examples of AI today are silicon based. For some humans, that may make the answer to Question 0 easy. Only humans can be persons; AI artifacts are not humans, so they should not be considered persons.

But many people think that the easy rejection of AI personhood is too glib. And once you reject the idea that only humans can be persons, the question becomes more nuanced. There are existing potential answers to this question in the literature. In a provocatively titled article, Joanna Bryson²¹ stakes out a strong position: “robots should be slaves.” Bryson’s abstract begins: “Robots should not be described as persons, nor given legal nor moral responsibility for their actions.” Bryson goes on to argue that we should not create machines that would seriously compete for that designation.

It seems a reasonable question to ask a corollary to Question 0: Why should we build machines that appear to be persons? Is it merely because we can? Is it because such machines give us functionality that is difficult or impossible

to get in other ways? These questions are, I think, important. They have been asked for years, but they are not as often answered, especially by the people and organizations developing and funding the development of these AI artifacts.

In the 10 years since Bryson’s publication, AI researchers have not appreciably slowed their work to make machines that increasingly closely resemble humans both in their behavior and their appearance. I anticipate that Question 0 and its corollaries will become far more contentious in the next few years. 

REFERENCES

1. N. Wiener, *Cybernetics or Control and Communication in the Animal and the Machine*. Cambridge, MA: MIT Press, 1948.
2. T. W. Bynum, “A very short history of computer ethics,” *APA Newslett. Philos. Comput.*, vol. 99, no. 2, p. 2, 2000.
3. N. Wiener, *The Human Use of Human Beings, Cybernetics and Society*. Boston: Houghton Mifflin, 1950.
4. A. Kaplan and M. Haenlein, “Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence,” *Bus. Horiz.*, vol. 62, no. 1, pp. 15–25, 2019. doi: 10.1016/j.bushor.2018.08.004.
5. J. McCarthy, M. Minsky, N. Rochester, and C. E. Shannon, “A proposal for the Dartmouth summer research project on artificial intelligence,” Stanford Univ., CA, 1955. [Online]. Available: <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>
6. T. W. Bynum, “Artificial intelligence, biology, and intentional states in computers and ethics,” *Metaphilosophy*, vol. 16, no. 4, pp. 355–377, 1985. doi: 10.1111/j.1467-9973.1985.tb00183.x
7. K. Čapek, R.U.R.: *Rossum’s Universal Robots* (P. Selver and N. Playfair, Transl.) 1920. [Online]. Available: preprints.readingroom.ms/RUR/rur.pdf
8. I. Asimov, “Runaround,” *Astounding Sci. Fiction*, vol. 29, no. 1, pp. 94–103, 1942.
9. “List of artificial intelligence films,” Wikipedia.org, 2020. [Online]. Available: https://en.wikipedia.org/wiki/List_of_artificial_intelligence_films
10. “List of fictional robots and androids: Television films and series,” Wikipedia.org, 2020. [Online]. Available: https://en.wikipedia.org/wiki/List_of_fictional_robots_and_androids#Television_films_and_series
11. K. Miller, M. Wolf, and F. Grodzinsky, “This ‘ethical trap’ is for roboticists, not robots: On the issue of artificial agent ethical decision-making,” *Sci. Eng. Ethics*, vol. 23, no. 2, pp. 389–401, 2016. doi: 10.1007/s11948-016-9785-y.
12. A. Howard and J. Borenstein, “The ugly truth about ourselves and our robot creations: The problem of bias and social inequity,” *Sci. Eng. Ethics*, vol. 24, no. 5, pp. 1521–1536, 2018. doi: 10.1007/s11948-017-9975-2.
13. R. Benjamin, *Race After Technology*. Polity, 2019.
14. M. Brundage et al., “The malicious use of artificial intelligence: Forecasting, prevention, and mitigation,” Future of Humanity Institute, Centre for the Study of Existential Risk, Center for a New American Security, Electronic Frontier Foundation, OpenAI. [Online]. Available: <http://arxiv.org/abs/1802.07228>
15. S. M. West, M. Whittaker, and K. Crawford, “Discriminating systems: Gender, race and power in AI,” AI Now Institute, New York, 2019. [Online]. Available: <https://ainowinstitute.org/discriminatingystems.html>
16. J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” *Proc. Mach. Learn. Res.* vol. 81, pp. 77–91, 2018. [Online]. Available: <http://proceedings.mlr.press/v81/buolamwini18a.html>
17. Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,”

- Science, vol. 366, no. 6464, pp. 447–453, 2019. doi: 10.1126/science.aax2342.
18. A. Kak, "The Global South is everywhere, but also always somewhere: National Policy Narratives and AI Justice," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc. (AIES '20)*, New York, 2000, pp. 307–312. doi: 10.1145/3375627.3375859.
 19. M. Wolf, K. Miller, and F. Grodzinsky, "Why we should have seen that coming: Comments on Microsoft's Tay "experiment," and wider implications," *ORBIT J.*, vol. 1, no. 2, pp. 1–12, 2017. doi: 10.29297/orbit.v1i2.49.
 20. F. S. Grodzinsky, K. Miller, and M. J. Wolf, "The ethics of designing artificial agents," *Ethics Inform. Technol.*, vol. 10, nos. 2–3, pp. 2–3, 2008. doi: 10.1007/s10676-008-9163-9.
 21. J. Bryson, "Robots should be slaves," in *Close Engagements With Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, Y. Wilks, Ed. Philadelphia: John Benjamins, pp. 63–74, 2010.

JASON BORENSTEIN is a director at the Graduate Research Ethics Programs, School of Public Policy and Office of Graduate Studies, Georgia Institute of Technology, Atlanta, Georgia, USA. Contact him at borenstein@gatech.edu.

FRANCES S. GRODZINSKY is a professor emerita at the School of Computer

Science and Engineering, Sacred Heart University, Fairfield, Connecticut, USA. Contact her at grodzinskyf@yahoo.com.

AYANNA HOWARD is a Linda J. and Mark C. Smith professor and chair at the School of Interactive Computing, Georgia Institute of Technology, Atlanta, Georgia, USA. Contact her at ah260@gatech.edu.

KEITH W. MILLER is a professor with the Department of Computer Science and the College of Education, University of Missouri–St. Louis, St. Louis, Missouri, USA. Contact him at millerkei@umsl.edu.

MARTY J. WOLF is a professor of computer science at Bemidji State University, Bemidji, Minnesota, USA. Contact him at Marty.Wolf@bemidjistate.edu.



IEEE TRANSACTIONS ON BIG DATA

▶ SUBSCRIBE AND SUBMIT

For more information on paper submission, featured articles, calls for papers, and subscription links visit: www.computer.org/tbd

TBD is financially cosponsored by IEEE Computer Society, IEEE Communications Society, IEEE Computational Intelligence Society, IEEE Sensors Council, IEEE Consumer Electronics Society, IEEE Signal Processing Society, IEEE Systems, Man & Cybernetics Society, IEEE Systems Council, and IEEE Vehicular Technology Society

TBD is technically cosponsored by IEEE Control Systems Society, IEEE Photonics Society, IEEE Engineering in Medicine & Biology Society, IEEE Power & Energy Society, and IEEE Biometrics Council

