Comment & Opinion

Ethical Algorithms: Promise, Pitfalls and a Path Forward

Jay J. Van Bavel
Associate Professor of Psychology & Neural Sciences
New York University
New York, New York
10003
jay.vanbavel@nyu.edu

Tessa V. West
Associate Professor of Psychology
New York University
New York, New York
10003
tessa.west@nyu.edu

Enrico Bertini
Associate Professor of Computer Science and Engineering
New York University
370 Jay Street
Brooklyn, NY 11201
enrico.bertini@nyu.edu

Julia Stoyanovich
Assistant Professor of Computer Science and Engineering & Data Science
New York University
370 Jay Street
Brooklyn, NY 11201
stoyanovich@nyu.edu

Word count = 1934 (2000 words)
Ethical Machines: Promise, Pitfalls and a Path Forward

The increasing role of machine-learning and algorithms in decision making has revolutionized areas ranging from the media to medicine to education to industry. As the recent One Hundred Year Study on Artificial Intelligence (Stone et al., 2016) reported: “AI technologies already pervade our lives. As they become a central force in society, the field is shifting from simply building systems that are intelligent to building intelligent systems that are human-aware and trustworthy.” Therefore, the effective development and widespread adoption of algorithms will hinge not only on the sophistication of engineers and computer scientists, but also on the expertise of behavioural scientists.

These algorithms hold enormous promise for solving complex problems, increasing efficiency, reducing bias, and even making decision-making transparent. However, the last few decades of behavioral science have established that humans hold a number of biases and shortcomings that impact virtually every sphere of human life (Banaji & Greenwald, 2013) and discrimination can become entrenched, amplified, or even obscured when decisions are implemented by algorithms (Kleinberg, Ludwig, Mullainathan, & Sunstein, 2018). While there has been a growing awareness that programmers and organizations should pay greater attention to discrimination and other ethical considerations (Dignum, 2018), very little behavioral research has directly examined these issues. In this paper, we describe how behavioral science will play a critical role in the development of ethical algorithms and outline a roadmap for behavioural scientists and computer scientists to ensure that these algorithms are as ethical as possible.

Earning trust in algorithm-assisted decision making is critical, since a lack of trust can harm individuals or groups by excluding them from access to resources, eroding their faith in institutions that employ algorithms, and reinforcing structural discrimination. Groups that distrust algorithms may be less likely to harness the potential benefits of this new technology. We argue that the existing literature on procedural justice, moral cognition, and social identity provides a powerful framework for guiding the development of human-algorithm interactions and offers a roadmap for research. It is unknown, for example, how different stakeholders perceive and evaluate the results produced by algorithms or what can be done to increase trust in algorithms. Yet this issue is at the heart of debates about the potential for bias in everything from Google searches to predictive policing to credit scoring.

Understanding why people see unfairness in these algorithms, and what can be done to increase the ethical and legal accountability of people and organizations using algorithms, will be critical for the sustained economic success of their use in business as well as the health of democratic institutions during a time of marked inequality and polarization (O’Neil, 2016). Organizations have a huge stake in understanding these issues since a failure to do so could result in bad press, angry consumers, market losses, and costly litigation. Thankfully, there is a rich tradition of research on these issues in the social sciences that offers useful insights into these contemporary
issues. We will discuss these issues and lay out a series of important, but unanswered, questions for behavioral scientists to explore during the age of algorithms.

**Procedural Justice**

Procedural justice is central to ensure that algorithms are fair and perceived as legitimate. Although this topic has been studied extensively in the domain of legal decision making, we believe it will have broad implications for algorithmic decisions. Procedural justice concerns the belief that as long as a process is fair, then people will accept outcomes that may not benefit them. Knowing what aspects of algorithmic decision-making people perceive to be fair and not fair is critical to understanding how they will react to outcomes. To achieve procedural justice, computer scientists and behavioral scientists will need to take different stakeholder's interest into account when building algorithms. Frameworks that accomplish this goal may increase feelings of procedural justice, as stakeholders feel like their perspectives influenced the development process (Lee et al., 2019).

Procedural justice plays an important role in a number of spheres. For example, employees who are faced with a pay cut are more likely to accept it if they believe the process of evaluating them was fair (Greenberg, 1990). Similar issues can be at play during hiring decisions or education admissions. For instance, students (and their families) may be more willing to accept a rejected admission to an elite school to the extent that they felt the process was fair and just. A recent lawsuit against Harvard University for bias in admissions decisions likely stems, in part, from a lack of transparency and procedural justice among some applicant groups. In fact, the lawsuit was filed by “Students for Fair Admissions”. Similar allegations of injustice were leveled against the New York City education department when it was revealed that only seven black students (out of 895 spots) had been admitted into New York’s most selective high school (Shapiro, 2019). In these cases, admissions decisions often hinge on algorithms implemented by humans (or a combination of humans and machines).

The role of procedural justice raises a number of important research questions. Do people trust algorithms more than a group of people making the same decisions (i.e., legitimacy)? Do people lose trust in algorithmic decision making when they learn that outcomes produce disparities? Is this only the case when these disparities harm their in-group? Which potential sources of bias are most likely to trigger trust or distrust in algorithms (e.g., lack of transparency, biased training dataset used to build algorithms)? These will be important research questions to tackle that also play a key role in maintaining perceived legitimacy in many organizations.

**Moral Cognition**

Moral cognition is concerned with how people determine whether an action or outcome is morally right or wrong. Decades of research on moral judgments have revealed that moral cognition is influenced by intuitions, and therefore is often inconsistent with reasoning and is applied unevenly (Haidt, 2007). Moral intuitions are often in conflict with rational decision making, making people reluctant to use machines that do not align with their values. There is
also a large body of evidence suggesting that people evaluate decisions made by humans differently from those made by computers (although this may be changing; see Bigman, Waytz, Alterovitz, & Gray, 2019) and people may be uncomfortable delegating certain types of decisions to algorithms. Understanding the nature of moral cognition is likely to play a big role in many domains of decision-making.

Recent work has found that moral cognition is central for programming driverless vehicles. Even though people approve of autonomous vehicles that might sacrifice passengers to save a larger number of non-passengers, they would prefer not to ride in such vehicles (Bonnefon, et al., 2016). Thus, utilitarian algorithms designed to minimize net harm may paradoxically increase casualties by decreasing the adoption of objectively safer technology. Failing to understand how people evaluate the moral programming of algorithms could thus undercut trust in certain algorithms and unwittingly cause harm to large groups of people and undercut a potentially massive market for driverless cars. The problem is compounded by the fact that moral preferences for driverless vehicles vary dramatically across cultures (Awad et al., 2018). Solving these sorts of problems will require an understanding of social dilemmas since self-interest might come directly in conflict with collective interest (Van Lange, Joireman, Parks, & Van Dijk, 2013).

There are several issues in the moral cognition literature relevant to the adoption and application of algorithms. Do people trust the moral cognition embedded within algorithms? Does this apply to some domains (e.g., pragmatic decisions, like clothes shopping) more than others (e.g., moral domains, like criminal justice sentencing)? Are people less likely to see disparate outcomes from algorithms as less morally repugnant (because they do not imply intent) or more repugnant (because they invoke taboo trade-offs)? Are there certain decisions that are simple taboo to delegate to algorithms (e.g., religious advice)? These are central elements of human moral cognition and will likely play a role in numerous algorithms. The issue of moral cognition also has clear implications for legal decision-making, since concepts like intent play a central role in sentencing.

Social Identity
Social identity is used to describe the subjective sense of self when people define themselves as part of a group or community (Tajfel, 1982). Decades of research on identity suggests that it plays a central role in our beliefs and interpretations of the world (Van Bavel & Periera, 2018). From an identity perspective, representation matters. People tend to trust in-group members and if they do not think their identity group is being represented in the decision making, they will not trust the system to make judgments that are in their best interest (see Alfano & Huits, 2019). This can explain why women and minorities are distrustful of hiring or admission decisions designed largely by white men, as well as why political parties are distrustful of algorithms developed by people with competing political identities. Of course, there are numerous other identities that may play a critical role in how algorithms are employed and evaluated.
The identity of the programmers as well as the stakeholders plays an important role in how algorithms are evaluated. One recent case where the role of identity has proven contentious is with search engines and news feeds. Liberal and conservative politicians are both demanding that big technology companies, like Facebook, become “Neutral” and they have repeatedly criticized Google for embedding bias in its algorithms. Even if such search algorithms prioritize high quality news sources over hyperpartisan news, partisans might believe that these algorithms are biased in favor of one political group over another if the results do not align with their political commitments. Even bad faith attacks on an algorithm can resonate among certain groups when issues of social identity are involved. In these cases, the identity of the programmers can overshadow more central features of the algorithm (e.g., the accuracy of the news source).

Research should examine how different group identities shape judgments of the outcomes of algorithms. Are certain communities less likely to trust algorithms due to the past mistreatment of their in-group? Are people more likely to see algorithms as biased against them if members of their own group were not involved in the construction or operation of the algorithms? Research suggests that people will trust their in-group more, especially when transparency is low and the process is ambiguous. As such, increased transparency and efforts to improve procedural justice may promote trust among a wider variety of identity groups. What kinds of transparency will promote trust among different stakeholder groups (and under what conditions will transparency increase distrust)? These are critical communities to promote fairness and trust in algorithms.

**Conclusion**

We argue that the marriage of expertise from the behavioral sciences and computer sciences is essential to develop algorithms that make ethical decisions and ensure the welfare of society. Insights from behavioral science could be used to generate questions that people can ask when considering implementing new algorithms (E.g., Have stakeholders been consulted about this algorithm? Do they feel it is fair? Is the training dataset representative of the population?) However, research on these questions is still in its infancy. Answering these questions will require expertise in psychology, politics, sociology, business, communications, philosophy, computer science, data science, and artificial intelligence, among other fields.

The future of these technologies has profound implications for society. This technology has already been applied to how we consume the news, where we work, what we eat, where we go to school, and even who we date. We offer a behavioral science framework on the promises and pitfalls of machine learning and artificial intelligence for fairness and equity in society. Understanding the factors that promote trust in machines is critical to ensuring that people engage fruitfully with algorithms. Applying the lessons from procedural justice, moral cognition, and social identity offers a powerful framework for harnessing this new technology.
References (max 15-20 references)


