A

AI and Corporate Responsibility

How and Why Firms Are Responsible for AI

Kirsten Martin[1] and Carolina Villegas-Galaviz[1,2]
[1]University of Notre Dame, Notre Dame, IN, USA
[2]Universidad Pontificia Comillas, Madrid, Comunidad de Madrid, Spain

Synonyms

AI; Artificial intelligence; Big data; Technology

Introduction

When companies develop and use technology, who is responsible for the moral implications design during development and impacting stakeholders during use can be contested. This entry explains how we think about corporate responsibility around the design, development, and use of AI.

AI and Corporate Responsibility

When a firm develops an AI program, that firm makes value-laden decisions as to who is important, who should be considered, and who can be ignored in a given decision. For example, in a mortgage approval program, the computer scientists train the algorithm on previous applicants including who was approved and rejected over a number of years. The AI program "learns" the attributes of individuals who are more likely to be approved. In any given data set, some people will be well-represented with all the data filled out and some will not have all their data included. Some types of people will be completely missing from the data. Data and computer scientists need to decide how much to punish people who are not represented or not well represented in the data. In addition, these same data and computer scientists make assumptions about missing data, how to treat outliers or edge cases, and how morally important it is to include more people in the model. In other words, if the predictive mortgage approval model does work well with certain people, should we care? Does it matter? How much should a bank care?

All this is to say that the firms that develop AI programs make value-laden decisions during design and development (Martin 2019). And that these decisions have moral implications for the firms that adopt the AI program and the users who are subject to a particular AI program. This runs counter to the mistaken belief that AI is somehow neutral or operates outside human involvement. In fact, these data and computer scientists have to make value-laden decisions throughout the development process.

D. C. Poff, A. C. Michalos (eds.), *Encyclopedia of Business and Professional Ethics*,
https://doi.org/10.1007/978-3-319-23514-1_1297-1

- *Training and Live Data*. When algorithms are developed from training data, who is represented in the training data and how the data is labeled directly impacts the creation of the algorithm. For example, when facial recognition is trained on primarily white men, the result is an algorithm who identifies white men moderately well but identifies black women incorrectly the majority of the time (Buolamwini and Gebru 2018). The model that is developed on a specific training data set may also be tailored to that training data and ineffective when applied to live data, causing harms, breaking rules, and reinforcing existing power dynamics.
- *Development of the Model*. Computer scientists make assumptions about the type of data, how the data is distributed, whether data is missing (and how bad is it for data to be missing), and whether the algorithm should care about outliers (and how much should it care). These are all value-laden decisions about individuals.
- *Outcome Chosen*. How does a particular outcome favor certain groups of people and how well does the outcome represent the phenomenon of interest? For example, we use GPA as a measurement for “good student in college” sometimes, but that does not mean that the GPA as an outcome is a good measurement of the phenomena we are interested in.
- *Mistakes*. All AI programs generate mistakes – people are mischaracterized and misidentified. Sometimes AI predicts someone will commit a crime and they do not, which is a false positive. Other times AI programs will predict someone will not commit a crime and they do which is a false negative. The types of mistakes (false positives versus false negatives) vary across decision contexts as well as which mistakes are more preferable for a given decision. For example, in the criminal justice system, we *prefer* false negatives: we prefer in the United States that someone be falsely set free rather than falsely imprisoned. Not only do computer scientists influence that types of mistakes that are more common with a given AI program, but they also influence whether or not the inevitable mistakes are able to be identified, judged, and fixed by users of the AI program. AI programs that are developed to be inscrutable, e.g., declared proprietary or designed to not be accessible by the firm that uses the AI program, allow the inevitable mistakes to continue by not allowing users to identify, judge, possibly fix mistakes
- *Contestability.* While people like to think that AI and related computer programing approaches are inscrutable, computer scientist Joshua Kroll notes that “inscrutability is not a result of technical complexity but rather of power dynamics in the choice of how to use those tools” (Kroll 2018). In other words, making a program difficult to use or making the mistakes created by the program difficult to identify, judge, and correct is a design decision. In fact, developers of AI programs should make their programs contestable (Mulligan et al. 2020), where subjects of the AI program are able to contest any decision made about them. This would require a certain amount of transparency and accountability in the design depending on the context of the decision and the types of users subjected to the program.
- *Assessment of AI*. The computer scientist influences how the AI program is assessed that it “works.” While we regularly, in the popular press and in academia, claim that AI is “accurate” or “efficient,” these measurements are actually constructed in the design for many programs. For example, one might need to know for whom is the program accurate and for whom is it not accurate. And, the efficiency gains for a company implementing AI programs may also mean that a bad decision is being made faster. We normally do not see mere efficiency as a goal for decision making. If we are hiring or arresting the wrong people, making those types of decisions faster with the aid of AI does not make the entire organization more efficient and may offload some of the work onto others. In fact, even the idea of prioritizing claims of accuracy and efficiency is a value-judgment that may work for the developing firm but not for the firm adopting and using the AI program.

Why Firms Are Responsible for AI

While the data and computer scientists make value-laden decisions in developing the AI program, the firm that *uses* the program is responsible for the ethical implications of their business decision. In other words, the bank is still responsible for making mortgage decisions, insurance companies are still responsible for adjudicating insurance claims, and firms are still responsible for their hiring decisions *even if* they augment their decision with an AI program. This places a distance between the moral decisions of development and the ethical implications in use.

Hence, the introduction of AI to decision-making increases what scholarship has called moral distance. Scholars use this concept to explain why individuals behave unethically towards those who are not seen. With AI decision-making, face-to-face interactions are minimized, and decisions are part of a more opaque process that humans do not always understand. Therefore, the issue regarding AI and moral distance is that firms miss the moral implications of their decision, for which they are responsible, being blinded behind the veil of AI (Villegas-Galaviz and Martin 2022).

Firms are responsible for the development, deployment, and use of AI in the same manner these same firms are responsible for the many business decisions they make about the products they develop, the materials they purchase, and the decisions about individuals that they make. Firms are responsible for the products and services they sell in that they have an obligation to not cause harm, act in a manner that does not further disadvantage the less fortunate, abide by the values and norms of society, and follow the law. Firms are similarly responsible for the decisions, augmented with AI, they make about individuals, employees, and users in that they have an obligation to treat people with dignity and respect, act as if individuals are an end and not a mere means to be used merely for the firms benefit, and to not create harm or diminish rights. The introduction of AI into an organization does not remove their responsibility for their actions.

Approaches to Take Responsibility for AI

Our ethical concepts, traditions, theories, and approaches can be seen as a way to close the gap between those making value-laden decisions and the ethical implications of those same decisions. In other words, these theories and approaches help the data and computer scientists understand better the ethical implications of their work. And, for firms adopting AI, these approaches provide a roadmap of the types of questions one should ask about the design and development and use of a specific AI program. Here we focus on more than mere consequentialism, which would only ask firms to calculate the possible net benefits or harms caused by the development, deployment, and use of AI. Consequentialism has the same deficits as an ethical tool when applied to AI decisions: the harms to the few who are considered marginalized, without a voice, or "edge" cases can be ignored in order to benefit the more powerful. Instead, we focus on those ethical approaches that would help firms take responsibility for AI and decrease the moral distance exacerbated by the use of AI.

1. Deontology

   In the field of AI and business ethics, much work has been done to find the right set of principles or AI ethical guidelines. Deontology, or principle-based ethics, bases the rightness of the action in that it follows the duty of those who act. Hence, individuals should decide according to their principles or rules rather than considering the consequences or context. These attempts within AI ethics usually follow the line to bring ethical frameworks from other disciplines, especially the four essential principles traditionally used in bioethics: beneficence, nonmaleficence, autonomy, and justice. However, scholars have brought out the fact that principles are not sufficient to guarantee ethical AI and the limitations of a principled approach to AI ethics (Mittelstadt 2019).
2. Justice and Fairness

   Fairness and AI has become almost synonymous with ethical AI, primarily when AI has been used to sort individuals, the program reinforces existing injustices captured in the data. For fairness and justice approaches to AI, initial works focused on how algorithmic decision-making processes do not lead to more objective and or more fair decisions than those by humans, who are seen as influenced by prejudice or emotions. In fact, AI has the potential to exacerbate issues regarding discrimination, bias, and fairness. In applying fairness approaches, best practice is to distinguish questions about discrimination and questions about justice. Fairness and justice theories highlight how being predicted or categorized should not be more likely for particular groups of people and that the system of allocating goods (the AI program) should not harm the less fortunate. Other approaches focus on equity, parity merit, and even the appropriateness of using particular attributes of individuals for a decision (Martin 2019). Discrimination law, on the other hand, focuses on ensuring the individuals are not treated or impacted differently based solely on a protected attribute (nationality, race, ethnicity, sexual orientation, gender, religion, etc.), and problems of discrimination are best examined throughout the process of design, development, and use of AI (Barocas and Selbst 2016).

3. Virtue Ethics

   Within virtue ethics approaches to understanding AI, the character traits of the agent or subject/user is the focus. Shannon Vallor's proposals lead the way to bring virtue ethics to answer the critical ethical questions of the current era. In her book *Technology and the Virtues* (2016), Vallor proposed a virtue-driven approach to the ethics of emerging technologies, such as AI, and a kind of ethical strategy for promoting the moral character needed for the challenges of recent times. In her framework, she adapted Aristotelian, Confucian, and Buddhist ethical reflections to create a set of what she calls are the *technomoral virtues* needed for the twenty-first century. The *technomoral virtues* framework is proposed to specify how humans should act to flourish in an uncertain future, where the uncertainty comes from the changing nature of emerging technologies. There, in search of a *technomoral wisdom*, the framework proposes an adaptation of 12 virtues to the new techno-social environment, in there are virtues like honesty, self-control, humility, or civility.

4. Ethics of Care

   More recently, the ethics of care has been used to better understand the moral implications of AI. The ethics of care is a contextualized moral theory focuses on interdependent relationships, individuals' vulnerabilities, circumstances, and the voice of the other in ethical decision-making. In AI ethics, the contribution of the ethics of care comes in line with the understanding of how AI models may marginalize those who do not fit within the pattern created and used by those who develop and deploy AI. In its critical aspects, the ethics of care can help in the comprehension of how algorithm decision-making can create harm and ignore the needs of individuals, especially the most marginalized groups (Villegas-Galaviz 2022).

5. Critical Approaches

   Critical theories attempt to understand the power dynamics and seeks to question not only the presumed objectivity and neutrality of analytics (Johnson n.d.) but also the power dynamics at play in building the algorithm, collecting and using the data, and deploying AI and analytics. Critical approaches seek to understand who gains and who is marginalized by the status quo. Langdon Winner (1980) is perhaps the most well-known scholar to take this approach to technology more broadly. Winner argues that technology, designed and used by society, has politics or "arrangements of power and authority in human associations." In regards to AI, critical approaches examine the development and use of AI through the lens of power – who retains power and who is marginalized – and usually makes the case for the lifting or emancipation of those who are being undermined by the use of AI.

Conclusion

When considering the ethical implications of development or use of AI to augment decisions, business practitioners and business ethics scholars have the tools to better understand how AI can be developed and used within the given values of the firm. AI does not fundamentally change how we think about ethics and responsibility.

Cross-References

- ▶ Artificial Intelligence and Business Ethics
- ▶ Artificial Intelligence and Ethical Journalism
- ▶ Artificial Intelligence and Teaching Values in Science
- ▶ Big Data Ethics
- ▶ Ethics and Artificial Intelligence

References

Barocas S, Selbst AD (2016) Big data's disparate impact. Calif Law Rev 104

Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: Proceedings of machine learning research, conference on fairness, accountability, and transparency, vol 81, pp 1–15

Johnson G (n.d.) Are algorithms value-free? Feminist theoretical virtues in machine learning. J Moral Philos

Kroll JA (2018) The fallacy of inscrutability. Philos Trans R Soc A Math Phys Eng Sci 376(2133):20180084

Martin K (2019) Ethical implications and accountability of algorithms. J Bus Ethics 160(4):835–850

Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intelle 1(11):501–507

Mulligan DK, Kluttz D, Kohli N (2020) Shaping our tools: contestability as a means to promote responsible algorithmic decision making in the professions. In: Werbach K (ed) After the digital tornado. Cambridge University Press

Vallor S (2016) Technology and the virtues: a philosophical guide to a future worth wanting. Oxford University Press

Villegas-Galaviz C (2022) Ethics of care as moral grounding for AI. In: Martin K (ed) Ethics of data and analytics. Taylor & Francis

Villegas-Galaviz C, Martin K (2022) Moral distance, AI, and the ethics of care. Available at SSRN: https://ssrn.com/abstract=4003468

Winner L (1980) Do artifacts have politics? Daedalus 109(1):121–136