MIS Quarterly Executive

Designing Ethical Algorithms

Algorithms drive critical decisions such as which patient is seen or who is offered insurance. Such algorithmic decisions, like all decisions, are biased and make mistakes. Yet, who is responsible for managing those mistakes? This article focuses on the responsibility of developers and users of algorithms to ensure algorithms support good decisions — including managing mistakes. First, while mistakes may be unintentional, ignoring or even fostering mistakes is unethical. Second, by creating inscrutable algorithms, which are difficult to understand or govern in use, developers may voluntarily take on accountability for the role of the algorithm in a decision.[1,2]

Kirsten Martin
George Washington University School of Business (U.S.)

Algorithms Raise Questions about Ethics and Accountability

Rapidly catching up to the growth of big data is the spread of advanced algorithms to make sense of these large datasets.[3] Algorithms are generally defined as a sequence of computational steps that transform inputs into outputs,[4] and range from simple if-then statements to artificial intelligence (AI), machine learning and neural networks. When applied to big data, algorithms create value from the digital data streams flowing through firms.[5] By 2020, predictive and prescriptive analytics will account for 40% of firms' net new investments in

1 Dorothy Leidner is the accepting senior editor for this article.

2 The author is thankful for the helpful guidance by Professor Leidner and the anonymous reviewers throughout the review process. The author is grateful for support from the National Science Foundation under Grant No. 1649415. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the author and do not necessarily reflect the views of the National Science Foundation.

3 Software to analyze big data is the second biggest driver of revenue within the IT industry, with sales expected to be more than $55 billion in 2019—more than twice that projected for hardware. For more information on the big data analytics and global IT markets, see: (1) Davis, J. "Big Data, Analytics Sales Will Reach $187 Billion By 2019," *Information Week*, April 24, 2016, available at https://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-sales-will-reach-$187-billion-by-2019/d/d-id/1325631; (2) Press, G. "6 Predictions For The $203 Billion Big Data Analytics Market," *Forbes*, January 20, 2017, available at https://www.forbes.com/sites/gilpress/2017/01/20/6-predictions-for-the-203-billion-big-data-analytics-market/; and (3) "Gartner Says Global IT Spending to Reach $3.7 Trillion in 2018," Gartner press release, January 16, 2018, available at https://www.gartner.com/newsroom/id/3845563.

4 Cormen, T. H., Leiserson, C. E., Rivest, R. L. and Stein, C. "Introduction to Algorithms", MIT Press, 2009.

5 "These applications typically employ advanced techniques, such as sophisticated algorithms, artificial intelligence and machine learning to splice, integrate and analyze real-time data, and to take decisions in real time in ways that can have a profound impact on creating business value." Quote from Anand, A. Sharma, R. and Coltman, T. "Four Steps to Realizing Business Value from Digital Data Streams," *MIS Quarterly Executive* (15:4), December 2016, pp. 250-277; see also: (1) Wixom, B., Yen, B. and Relich, M. "Maximizing Value from Business Analytics," *MIS Quarterly Executive* (12:2), June 2013, pp. 37-49; and (2) Pigni, F, Piccoli, G. and Watson, R. "Digital Data Streams: Creating Value from the Real-Time Flow of Big Data," *California Management Review* (58:3), May 2016, pp. 5-25.

SIM WHERE IT LEADERS CONNECT
KELLEY SCHOOL OF BUSINESS INDIANA UNIVERSITY

business intelligence and analytics for tasks such as tagging, categorization, clustering, question answering, filtering, and alerting.[6]

To date, the focus of research has been on the use of increasingly complex algorithms to create value for a firm and enhance customer service.[7] Yet, problems with algorithmic decisions increasingly reach the press with headlines such as "What happens when an algorithm cuts your health care?" or "How to persuade a robot that you should get the job." While big data has received its fair share of criticism,[8] now algorithms are scrutinized as being unfair, inscrutable, causing harm and diminishing rights. Researchers ask if data scientists should take a Hippocratic oath and call for algorithms to learn without prejudice.[9] Computer scientists meet to understand fairness, accountability, and transparency in machine learning.[10] Firms now need to understand not only how to create value in the design, development and use of AI but also answer questions about the governance of such algorithmic decisions.

Two facets of more complex and autonomous learning algorithms force questions about ethics and accountability into the conversation. First, AI has become ubiquitous and cheap, therefore pushing algorithmic decisions throughout the organization, including decisions that are customer-facing.[11] Such "edge" technology means decisions and mistakes are felt by outsiders and identified by researchers and the press. In other words, algorithmic decisions are being noticed and reported. Second, algorithmic decisions are faster and include less human analysis; organizations are not taking time to consider how AI can and should displace the judgment of individuals and how algorithmic decisions should be governed, as they did when large ERP systems were first implemented.

What does this mean for IS and computer science professionals and their responsibility for developing an algorithm to sell or choosing an algorithm to use? Here, I focus on algorithms as active, opinionated participants in algorithmic decisions, which, like all decisions, make mistakes. I leverage what we know about effective decision-making in firms to highlight the types of mistakes we can expect from algorithms and how to better identify, judge and correct those inevitable mistakes. In effect, all algorithmic decisions will produce mistakes; but ethical algorithms will offer a mechanism to identify, judge and correct mistakes. I offer two mechanisms from ethical decision-making—social embeddedness and reflection—as tools for designing an algorithm for greater individual accountability within a business decision. Here, the onus shifts to the algorithm's developer to design who is responsible for identifying mistakes, judging mistakes as appropriate (or not), and correcting those mistakes.

I suggest we categorize algorithms not based on the technical specifications (such as linear programming or machine learning) or the type of task performed (such as to categorize, describe, prescribe, sort etc.), but rather based on the degree to which the algorithm is designed to be inscrutable and take on a larger role in a decision.[12] Importantly, by creating inscrutable, autonomous algorithms, firms may voluntarily take on accountability for the role of the algorithm in the decision, including the ability to govern the inevitable mistakes.

The Role of Algorithms in Decision-Making

Algorithms, including AI, machine learning, and neural networks, are designed to take on the work of individuals within decisions. When an algorithm edges out individuals from performing tasks in a decision, then these roles

6 Columbus, L. "Roundup of Analytics, Big Data & BI Forecasts And Market Estimates," 2016, *Forbes*, August 20, 2016, available at https://www.forbes.com/sites/louiscolumbus/2016/08/20/roundup-of-analytics-big-data-bi-forecasts-and-market-estimates-2016/.

7 Watson, H. J. "Preparing for the Cognitive Generation of Decision Support," *MIS Quarterly Executive*, (16:3), September 2017, pp. 153-169.

8 Martin, K. E. "Ethical Issues in the Big Data Industry," *MIS Quarterly Executive* (14:2), June 2015, pp. 67-85.

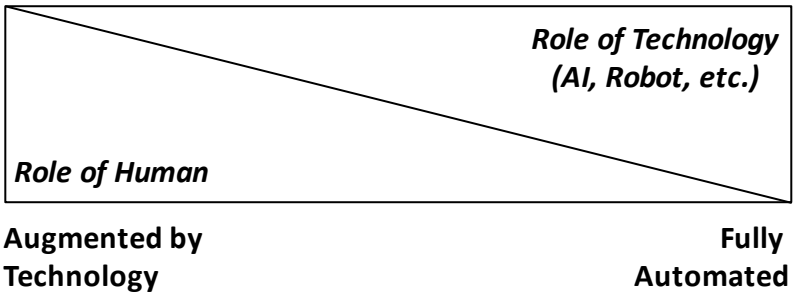
9 Lewis, H. "In 2018, machines must start to learn without prejudice," *WIRED*, January 3, 2018, available at http://www.wired.co.uk/article/technology-prejudice-artificial-intelligence-helen-lewis.

10 For example, the *ACM Conference on Fairness, Accountability, and Transparency* (ACM FAT*) brings together researchers and practitioners interested in fairness, accountability and transparency in socio-technical systems. For more information, see https://fatconference.org.

11 Ives, B., Palese, P. and Rodriguez, J. A. "Enhancing Customer Service through the Internet of Things and Digital Data Streams," *MIS Quarterly Executive* (15:4), December 2016, pp. 279-297.

12 Diakopoulos, N. "Accountability in Algorithmic Decision Making," *Communications of the ACM* (59:2), February 2016, pp. 56-62. Diakopoulos rightly identifies the important roles of algorithms in prioritizing, classifying, associating, and filtering individuals

Figure 1: The Relative Roles of Individuals and Technology in Decision-Making



and responsibilities do not disappear. Algorithms relieve individuals from the burden of certain tasks, similar to how robots edge out workers in an assembly line. Similarly, algorithms are designed for a specific point on the augmented-automated continuum of decision-making in Figure 1. In choosing a point along this continuum, developers, make a moral choice as to the delegation of tasks and responsibilities between algorithms and individuals within decision systems.

Less discussed is how an algorithm can influence the delegation of who-does-what within a decision. At a minimum, technologies alleviate the need for others to do a task.[13] Algorithms can also suggest that others perform tasks or even preclude individuals from an important task. For example, an algorithm assumes data is in a particular format and assumes someone will provide (and clean) that data. By making an algorithm proprietary, an algorithm can preclude a court from offering due process rights to defendants or prevent a Facebook user from identifying the source of a news story appearing in a newsfeed.

Fortunately, we have encountered this range of roles and the associated questions about responsibility with robots[14] and automation, including questions about sharing moral responsibility with robots and integrating ethics in design for engineers.[15] From the perspective of human factors engineering (i.e., those who study the automation of processes), the most important question in design is the division of labor between robots and humans depicted in Figure 1, because design becomes hard to change once technology is in use.

While allocating tasks and responsibilities between individuals and technology is not new, with algorithms, this delegation is happening faster and in a new area of decision-making. Complicating the analysis is the mistaken perception that algorithmic decisions are objective, when algorithms are actually quite value-laden and are designed for a preferred set of actions and view of how the world will and should work.[16] Just as robots are analyzed as members of an assembly line—and must support the rules and norms of manufacturing—so too algorithms should be analyzed as actors within a decision.

13 See Latour, B. "Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts," in *Shaping Technology/Building Society: Studies in Sociotechnical Change,* W. Bijker, and J. Law, MIT Press, 1992, pp. 225-58. Latour uses physicists looking for "missing mass" in the universe as a metaphor for sociologists or ethicists looking for missing responsibility in a system of technologies and individuals.

14 Dodig-Crnkovic, G. and Persson, P. "Sharing Moral Responsibility with Robots: A Pragmatic Approach," *Proceedings of the 2008 conference on Tenth Scandinavian Conference on Artificial Intelligence: SCAI 2008*, 2008, pp. 165-168.

15 Cummings, M. L. "Integrating Ethics in Design through the Value-Sensitive Design Approach," *Science and Engineering Ethics* (12:4), December 2006, pp. 701-15; see also: (1) Hellström,T. "On the Moral Responsibility of Military Robots," *Ethics and Information Technology* (15:2), June 2013, pp. 99-107; and (2) Lokhorst, G.-J. and van den Hoven, J. "Responsibility for Military Robots," in *Robot Ethics: The Ethical and Social Implications of Robotics* (P. Lin, K. Abney, and G. A. Bekey), The MIT Press, 2011, pp. 145-156. These two articles address the ethics of military robots as well as the responsibility when robots kill.

16 Martin, K. E. "Ethical Implications and Accountability of Algorithms," *Journal of Business Ethics*, June 2018, pp. 1-16. This article conceptualizes algorithms as value-laden by (1) creating moral consequences, (2) enabling and diminishing stakeholder rights and dignity, and (3) reinforcing or undercutting ethical principles.

Algorithmic Decisions Make Mistakes

Framing algorithms as taking on a role within a decision changes how we think about designing algorithms, because an important task in decisions concerns mistakes. All decisions contain the possibility of mistakes, and better decisions contain a vehicle to identify, judge, and fix mistakes. In manufacturing, the decision to ship final inventory includes a check to identify flaws, judge if the flaws are within an error range, and (if needed) assign someone to fix the mistake. Alternatively, a machine could be designed to ship inventory without allowing for any of these steps, thereby precluding humans from identifying, judging, and correcting mistakes. For example, the shipping label could be glued on the final product and shipped directly from the machine.

In general, managers, firms, and management researchers persistently seek to understand bad business decisions and avoid mistakes. Decisions can be unethical, unfair, bad for the long-term value creation for stakeholders, or just self-defeating. Firms and managers make bad decisions due to bad inputs (myopic, limited sources), bad reasoning (maximizing on a single objective function) and bad execution (sloppiness, laziness, lack of courage). In doing so, managers regularly do things they should not such as promote the wrong person, and not do things they should, such as pass over a good hire. Management scholars research how to minimize and manage these bad decisions. The goal is to support good decisions, that create value and minimize mistakes.

Algorithmic decisions are no different. Algorithms, whether as merely augmenting or automating human decisions, are used in important organizational decisions such as who is hired, who is fired, whether someone is deemed a terrorist, the terms offered for financing, whether an insurance company negotiates over a claim, and even how someone is sentenced.[17] In other words, we need to ask whether and how algorithmic decisions produce biased "answers" or mistakes, to categorize the mistakes and discuss who should be responsible for managing the mistakes. These mistakes destroy value, lead to bad decisions, and end up on the front page of the newspaper.

Mistakes—an action or judgment that is misguided or wrong—need not necessarily be unethical or unfair. Mistakes occur all the time in business and in life due to mistaken information or reasoning. However, ungoverned decisions, where mistakes are unaddressed, nurtured, or even exacerbated, are unethical. Ungoverned decisions show a certain casual disregard as to the (perhaps) unintended harms of the decisions; for important decisions, this could mean issues of unfairness or diminished rights. Further, some algorithmic decisions learn from previous decisions and can therefore quickly cause mistakes to impact thousands if not millions of decisions. In other words, while mistakes may be inadvertent, governance decisions are not. A lack of intentionality may be a fair excuse for a mistake but not a valid excuse for not governing mistakes.

Below I explain how algorithms may be designed to preclude individuals from identifying, judging and correcting mistakes and, therefore, take on the responsibility for those mistakes.

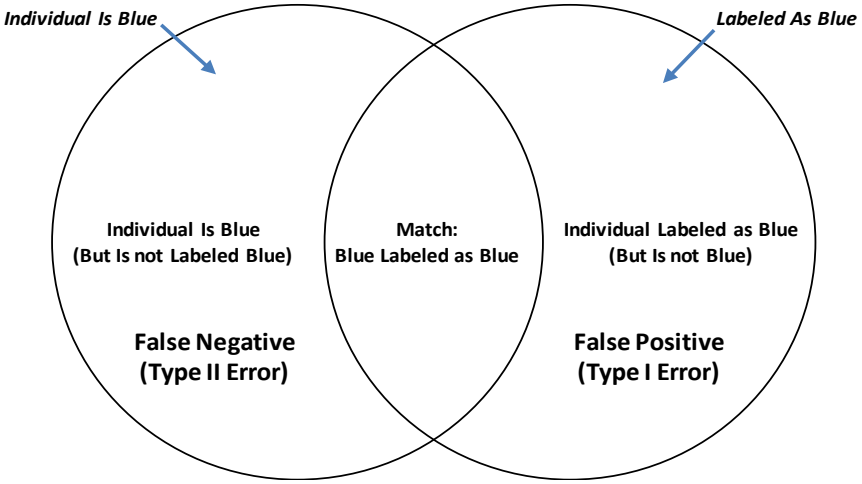
Identifying Mistakes in Algorithmic Decisions

The Algorithmic decision mistakes fall into two classes—category mistakes and process mistakes.

Category Mistakes. Algorithms that categorize and prioritize individuals, such as individuals who need an ad, prefer a search result, are employable, are a terrorist, have cancer, etc., scan large datasets to label individuals. These algorithmic decisions are vulnerable to two types of classic mistakes, which

17 See: (1) Angwin, J., Kirchner, L., Larson, J. and Mattu, S. "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's biased against blacks," *ProPublica*, May 23, 2016, available at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; (2) Brown, K. When Facebook decides who's a terrorist, *Splinter*, October 11, 2016, available at http://fusion.net/story/356354/facebook-kashmir-terrorism/; and (3) O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown Publishing Group, 2016.

Figure 2: Types of Mistakes in Algorithmic Decisions



I call category mistakes. First, false positives, or Type I errors, are the incorrect assignment of a label. For example, when someone is labeled as a terrorist when they are not, when someone is categorized as having cancer when they do not, or when someone is labeled as a future criminal when they are not. False positives are when the algorithmic decision (or human-centric decision) scans the universe of individuals and mistakenly labels the individual as within the preferred category.

Alternatively, false negatives, or Type II errors, incorrectly exclude someone from a category; false negatives encompass letting someone slip away by not labeling them. For example, identifying someone as not a terrorist when they are, categorizing someone as not employable when they are, or labeling someone as not a future criminal when they are. False negatives entail the algorithmic decision scanning the universe of individuals and not labeling them as the preferred category when the label may actual fit. Figure 2 illustrates the types of mistakes decisions.

Importantly, all decisions, both human-centric and algorithmic, contain a probability for each type of mistake. And, the likelihood of each type of mistake is not necessarily symmetrical, in general or across specific groups of individuals, as depicted in Figure 2. The mistake could be more frequently found in one group of individuals, making the mistake itself biased. For example, recent work in facial recognition illustrates that the distribution of category mistakes is not consistent across races and ethnicities: facial recognition algorithms are good at identifying white males and regularly misidentify black females.[18]

Process Mistakes. In addition to categorizing incorrectly, algorithms can make mistakes in the process of making the decision. Whereas category mistakes show up in the outcome of the algorithm, process mistakes occur when an algorithm makes a mistake in how the decision was made, regardless of the outcome. Table 1 compares these types of mistakes for different decision contexts such as education, public policy, health care, banking etc. Each context has norms as to the type of factors that should be considered in making a decision. When a doctor is making a diagnosis and treatment plan, using your friends' high school GPA[19] would be inappropriate and outside the norm of the decision. Similarly, when being approved for public housing or food stamps, considering the applicant's father's undergraduate degree would be inappropriate. Particularly with machine learning or neural networks (i.e., algorithms that "learn" what factors are important from existing data), the resulting decision may inadvertently use inappropriate factors in the decision—even

18 Buolamwini, J. and Gebru, G. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (81), 2018, pp. 77-91.

19 Grade point average—a number representing the average value of the accumulated final grades earned in courses over time.

Table 1: Types of Mistakes in Algorithmic Decisions

Mistake\ Context	False positive: Incorrectly include in a category	False negative: Incorrectly exclude from a category	Process Mistake: What factors drive the algorithmic decision? How is data gathered and used?
Manufacturing	Shipping a product as finished when it is actually defective	Rejecting a perfectly good product as defective	Deciding to ship a product only because it will help hit sales targets
Contacts and Friends	Identifying someone as a friend who is not	Not listing a friend (who may be a great fit)	Identifying friends based on individuals attending AA meetings
Political/ Advertising	Placing the wrong ad	Not placing the right ad	Targeting ads based on a medical condition; Google following users to see if advertising works
Social Services/ Public Goods	Family is given access to food stamps or Medicaid when they do not qualify	Family services program failing to flag toddlers who are in danger	Considering race when determining how to allocate police in a city
Judicial	Incorrectly labeling someone as a future criminal	Labeling someone as not a future criminal when they are	Considering a defendant's father's criminal history in categorizing risk of re-offending
Housing	Approving housing application for someone who doesn't qualify	Denying someone housing who does qualify	Placing a housing-related Facebook ad that excludes blacks, Asians, and Hispanics
Employment	Promoting the wrong person	Rejecting a good candidate	Considering a candidate's marital status
Location	Categorizing someone as at home when they are not	Deciding someone is not at a store when they are	Strava's heatmap software identifying U.S. military bases overseas

when not designed to do so. Previous work has highlighted how algorithms must abide by procedural norms, including considerations of due process, disparate treatment and impact, and norms of justice.[20] These types of process mistakes may be by design or learned by the algorithm from biased training data.

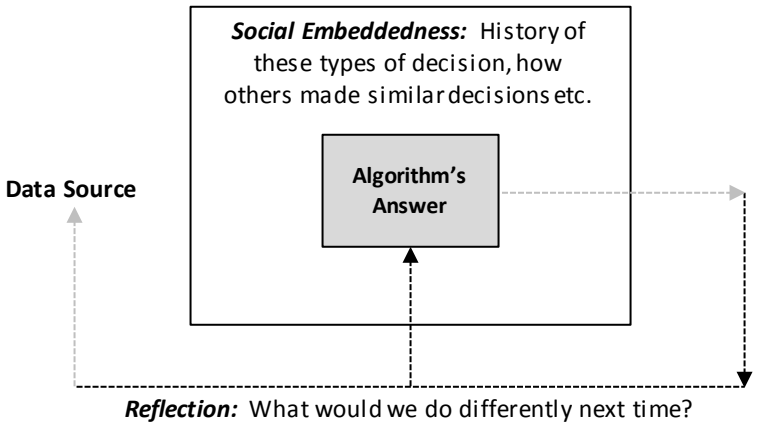
Judging Mistakes

Within a given context, certain types of mistakes are preferred, and not all mistakes are a cause for concern. For a medical decision, the preference may be to mistakenly identify someone as having cancer rather than letting cancer go undetected. The medical community tends to avoid false negatives, whereby a patient is not labeled as sick when the individual is actually sick or hurt. The justice system tends to avoid false positives for convictions and has a slight preference to not mistakenly find someone guilty who is actually innocent. However, the COMPAS algorithm is an interesting example: black defendants were more likely to be mistakenly labeled "likely to re-offend" (when they were not), compared to white defendants.

Even within a type of decision, mistake preferences are not necessarily consistent. Firms may find nothing worse than hiring the wrong person or categorizing someone as a "good hire" (when they are not), thereby avoiding a false positive. However, earlier in the hiring process, greater diversity can be achieved by being overly inclusive in who is brought in for an interview. The preference earlier in the hiring process is to label someone as possibly good even if they are not, thereby preferring a false positive. Even within hiring, the preferred type of mistake may shift. Importantly, the appropriateness of

20 For more information, see: (1) O'Neil, C., op. cit., 2016; (2) Citron, D. K. "Technological Due Process," *Washington University Law Review* (85:6), 2008, pp. 1249-; and (3) Barocas, S. and Selbst, A. D. "Big Data's Disparate Impact," *California Law Review* (104:3), June 2016, pp. 671-732.

Figure 3: Adding Social Embeddedness and Reflection to Algorithmic Decisions



a mistake, the risk tolerance and error ranges for mistakes, and the preference for a type of mistake is contingent upon the decision context and would need to be considered in the design and use of the algorithm. Mistakes occur in all decisions, and certain types of mistakes are preferred depending on the context of the decision.

Correcting Mistakes

Finally, algorithmic decisions need an ability to correct mistakes by adjusting the algorithm's outcome in the larger decision—particularly when the outcome feeds back into the dataset used to train or test the algorithm. Machine learning algorithms learn from existing data what factors are important for a given result. If uncorrected, the mistakes can feed into a cycle whereby the mistake becomes a part of the dataset the algorithm depends upon. And, when an algorithm creates mistakes with increasing frequency, the technology appears to learn from current mistakes, create "answers" that are mistakes, and contribute to a new data set that is riddled with mistakes from which future algorithms will learn—thus creating a biased cycle of discrimination with little human intervention required.[21]

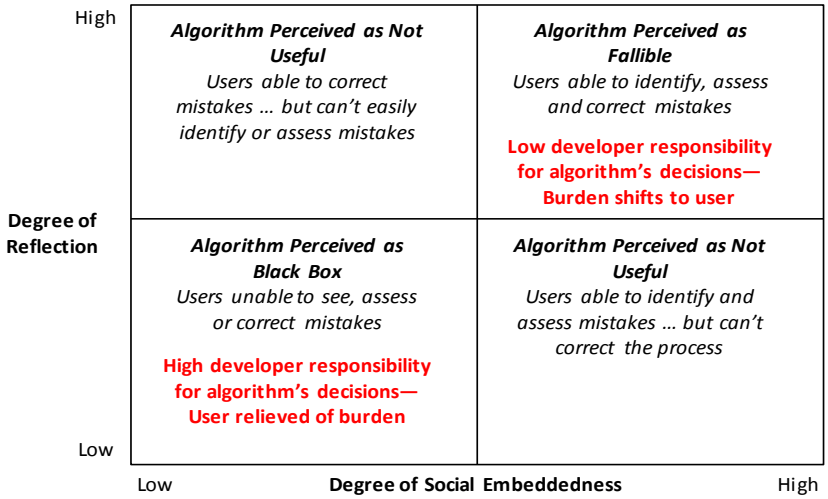
Algorithms and Ethical Decision-Making

Mistakes Go Unnoticed in the Current Algorithmic Decision Model

Mistakes can easily be missed due to the current model of algorithmic decision-making that presumes a rational decision model with linear processing and a goal of "efficiency." Mistakes can be missed because of an algorithm's artificially inflated role within a decision, where the algorithm is framed as a powerful yet inscrutable entity that does not make mistakes. Algorithms can wrongly be presumed to be clean or not biased and viewed with a veneer of objectivity, where individuals defer to the perceived power of the very notion of an algorithm. In addition, algorithms are a less visible part of the decision and often less accessible to question—even being held secret. The current approach to algorithmic decision-making runs the danger of treating the algorithmic process and output as both inevitable and final, where the algorithmic outcome cannot be questioned or changed, and mistakes are left ungoverned. Fortunately, decision-making scholarship offers solutions to both of these objectivity problems.

21 Cathy O'Neil refers to these types of exacerbating impacts—where the algorithm produces biased mistakes, impacts the less fortunate and does so at the velocity associated with big data initiatives—as weapons of math destruction. See O'Neil, C., op. cit., 2016.

Figure 4: Social Embeddedness and Reflection Combine to Shift Accountability for Decisions



Adding Social Embeddedness Helps to Identify and Judge Mistakes

The problem of viewing a decision as inevitable can be countered by acknowledging the context, or social embeddedness,[22] of the algorithmic decision-making process: how the algorithmic process and output could have been done differently and produced a different outcome. In more human-centric decision-making, social anchoring helps put the decision into context and perspective by checking in with others. Similarly, philosopher Richard Rorty calls for greater contingency to put quandaries into perspective.[23] For algorithmic decisions, algorithm developers could add visualization to show how the output such as a defendant's risk assessment score, compares to others committing the same crime or to those from the same state, illustrates how sensitive the outcome is based on the assumptions made, allows the user of the algorithm to change some of the input variables to see how the answer changes, or provides sensitivity tests. Such a contingent approach would be part of the design and development of the algorithm. Importantly, mistakes can only be identified if the output is placed into the perspective of similar decisions, the larger context, and historical decisions as depicted in Figure 3. The historical perspective may not necessarily be better or more desirable, but it does offer a way to measure progress in striving for a better decision.

Adding Reflection Helps to Correct Mistakes

Second, the issue of viewing algorithmic decisions with a degree of finality suggests users do not question changes for the future, as if the algorithm and the surrounding decision-making assemblage offer the best we have to offer without mistakes. In human-centric decision-making, reflection in decisions calls for the ability to go back to revisit, challenge and question the outcome and process; in pragmatic terms, Richard Rorty calls on us to not treat the decision like a final vocabulary but rather with an ironic view of the decision.[24] For algorithmic decision-making, designers would need to inscribe[25] the

22 Martin, K. E. and Parmar, L. P. "Assumptions in Decision Making Scholarship: Implications for Business Ethics Research," *Journal of Business Ethics* (105:3), May 2012, pp. 289-306.

23 Rorty, R. "Contingency, Irony, and Solidarity," *Cambridge University Press*, 1989; see also Sonenshein, "The Role of Construction, Intuition, and Justification in Responding to Ethical Issues at Work: The Sensemaking-Intuition Model," *Academy of Management Review* (32:4), October 2007, pp. 1022-1040.

24 Rorty, R. op. cit., 1989.

25 Madeleine Akrich argues that "… a large part of the work of innovators is that of 'inscribing' this vision of (or prediction about) the world in the technical content of the new object." Designers of technology—including algorithms—make assumptions about what the world will do and inscribe during design how their technology will fit into that world. See Akrich, M. "The De-Scription of Technological Objects," in *Shaping Technology/Building Society: Studies in Sociotechnical Change,* W. Bijkerand J. Law (eds.), MIT Press, 1992, pp. 205-224.

ability to go back to question algorithmic output with due process and reflection. For example, in using algorithms for worker evaluations, such as analyzing technology workers for idea generation or sifting through potential employees for a job, a weakness in judging the effectiveness of the algorithm is the difficulty of finding false negatives—i.e., people the algorithm falsely labels as "bad." The company does not know what happened to the good applicant that got away and therefore how ineffective the algorithm might be. However, such examinations are possible. As noted by Cathy O'Neil, the author of *Weapons of Math Destruction*, Amazon goes to great lengths to make sure the "right" decision is made, in terms of customer retention and marketing techniques, and is able to find the false negatives and correct the algorithm, illustrating that reflection is possible if designed into the algorithm. Algorithmic decision-making can incorporate the ability to revisit the answers to ensure that the classification is working as desired and not creating mistakes.

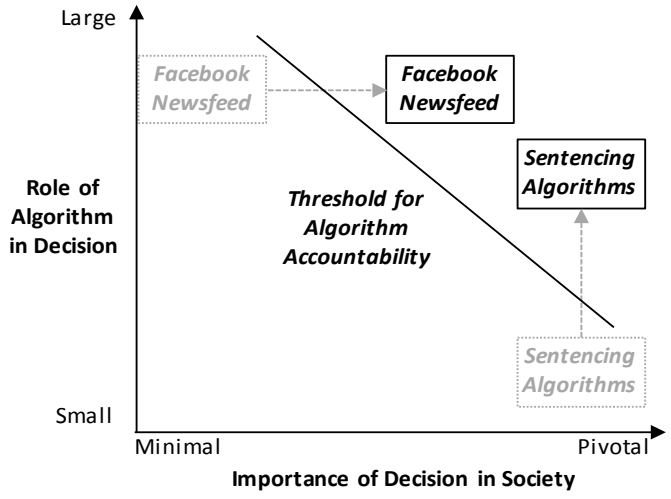
Designing Accountability for Mistakes into The Algorithm

Thus far I have argued that algorithmic decisions include mistakes like all decisions, and better algorithmic decisions account for who is responsible to identify, judge, and correct mistakes. In addition, this delegation is done in design: developers of algorithms inscribe their vision of who will be responsible for mistakes through the degree of social embeddedness and reflection permitted in use. Both social embeddedness and reflection work to allow users greater accountability to identify, judge, and correct mistakes, as shown in Figure 4. In other words, in designing social embeddedness and reflection into the algorithmic decision, developers of the algorithm permit users to take responsibility for governing the algorithmic decision.

However, how much accountability or how large a role should users versus algorithms have in the decision? Where along the augmenting-automation continuum of Figure 1 should we design the algorithm? One possibility is assessing the appropriate role and associated responsibility attributed to an algorithm as contingent upon the type of decision being made. Here, I argue the limits of the algorithmic accountability in the decision is dependent upon the type of decision. Accountability for the algorithm in the decision is reframed as a design decision; and the appropriate role of the algorithm in the decision is based on the relative importance of the decision in society.

Figure 5 illustrates one example of a threshold model of algorithmic accountability where the algorithm would be categorized by the role of the algorithm in a decision (y-axis) and by the importance of the decision in society (x-axis). For example, we regularly give extra scrutiny to decisions about the delegation of social goods

Figure 5: Threshold Model for Algorithm Accountability



(education, health care), the acknowledgement of rights (imprisonment, safety), and critical moments (credit decisions, buying a house), but are less concerned about deciding the color of the paint of the roads or placement of an advertisement for cereal.[26] We can think of decisions as falling along a range of importance within society. Not all decisions warrant equal scrutiny, with some having minimal importance and others being pivotal in the lives of individuals and society.

In addition, for the y-axis, algorithms can be designed to take on a role within a decision as per Figure 1. Algorithms with a greater role in a decision preclude individuals from governing the decision process, whereas algorithms with a smaller role offer greater social embeddedness and reflection for users. Importantly, this categorization is irrespective of the technical specifications of the algorithm.

The issue of how to allocate accountability between technology and individuals is not new. When designing autopilots for aircraft, we purposefully delegate roles and responsibilities to humans to create what are referred to as "moral crumple zones" where the human bears the brunt of the moral penalties when the overall system fails—not because the human is required but because the decision is too important to let the computer program decide autonomously.[27] Similarly, the goal of military development of technology has moved away from increasing automation to more of a focus on "robots supporting human decision making."[28] These examples acknowledge both the need for individuals to have a larger role in algorithmic decisions in some cases and that the role of the algorithm in a decision is constructed in design.[29]

Based on the arguments above, the appropriate role of an algorithm in a decision may be inversely proportional to the importance of the decision in society: the more important the decision, the more we expect a human agent to take responsibility within the decision with greater social embeddedness and reflection. For example, when making medical decisions, doctors may refer to IBM's Watson program to augment a medical diagnosis but still make the decision themselves.[30] The people who develop Watson should design the algorithm with the appropriate rules about mistakes in mind for that given decision—and be held accountable for those mistakes if they choose to preclude humans from identifying, judging, and correcting mistakes by making the algorithm inscrutable. Currently, Watson is designed to allow individuals to make the final decision and decide if the categorization is correct, as is appropriate for a medical diagnosis.

Two examples further illustrate this point: changing the role of the algorithm in a decision and examining the change in the decision's importance in society.

Changing the Role of the Algorithm in a Decision

The decision to sentence a defendant, or the decision to take away an individual's rights for a set amount for time, is widely understood as pivotal; we have laws stipulating how this decision should be made. Recently, some U.S. courts have been using a risk assessment algorithm, COMPAS, to guide how to sentence defendants. However, what factors drove

26 In relation to categorizing the societal importance of decisions, Zynep Tufekci refers to the importance of gatekeeping algorithms used as subjective decision makers, whereas Ryan Calo and Jenna Burell focus on consequential decision-making, and Cathy O'Neill refers to pivotal decisions in someone's life as deserving more attention. Each acknowledges that not all algorithm-based decisions are of equal moral importance. Cathy O'Neil's term—pivotal decisions—for the role of decisions in society has been used for the x-axis in Figure 5. For more information, see: (1) Tufekci, Z. "Algorithmic Harms beyond Facebook and Google: Emergent Challenges of Computational Agency," *Colorado Technology Law Journal*, (13), 2015, pp. 203-216; (2) Calo, R. "Artificial Intelligence Policy: A Roadmap," *SSRN*, August 8, 2017, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3015350; (3) Burrell, J. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms," *Big Data & Society* (3:1), January 2016; and (4) O'Neil, C. op. cit., 2016.

27 Elish, M. C. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction," *Engaging Science, Technology, and Society* (5), 2019, pp. 40-60, available at https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2757236.

28 Johnson, D. G. "Technology with No Human Responsibility?," *Journal of Business Ethics* (127:4), April 2014, pp. 707-715.

29 Meg Jones refers to this concept (the need for an individual to have a larger role in a decision) as "a right to a human in the loop that is intended to protect the dignity of the data subject," and, I would argue, for pivotal decisions. See Jones, M. L. "The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood," *Social Studies of Science* (47:2), April 2017, pp. 216-39.

30 Marks, N. Rawaf, A. and St. John, M. "Artificial Intelligence Positioned to Be a Game-Changer," CBS News, *60 Minutes*, June 25, 2017, available at http://www.cbsnews.com/news/artificial-intelligence-positioned-to-be-a-game-changer/.

the decisions within the algorithm was not available for defendants to question even when the outcomes appeared to be biased based on race; further, the outcomes were biased. For the COMPAS case of an algorithm used for sentencing, the increasing role of the risk assessment algorithm became problematic for such an important decision as depicted in Figure 5. In other words, the algorithm took on too large a role in the decision (y-axis) for this type of societal decision (x-axis)—particularly since social embeddedness and reflection were not designed into the algorithm.

A similar example is the use of an algorithmically curated dossier used to weed through applicants for jobs.[31] In the example, Catherine Taylor was denied a job at the Red Cross due to errors in how she was identified; the algorithm picked up damaging facts about the wrong person and attributed them to Ms. Taylor. Later, based on this false attribution, Ms. Taylor's application for federal housing was also rejected. However, this time, an official assumed a larger role concerning Ms. Taylor's housing application and questioned the veracity of the algorithm's identification matches. In effect, the federal housing official did not take the algorithm's answer as final and took on a larger role in the decision, thereby diminishing the role of the algorithm—as would be appropriate for a decision allocating a social good. Accountability was shifted to the right of the threshold line in Figure 5.

Importantly, the degree to which an algorithm is inscrutable contributes to our ability to identify, judge, and correct mistakes in algorithmic decisions. An algorithm's opacity, or degree of inscrutability, may be purposeful due to corporate secrecy or deception (see also Pasquale) but can also be due to the specialized skill required that is not currently understood, and to the challenges of scale and complexity of machine learning algorithms.[32] Opacity, however, need not necessarily be due to the inscrutability of machine learning or neural networks, and stakeholders to the decision need to push back on the immediate response of "it's complicated" when companies are asked how algorithmic decisions are made.[33]

Examining the Role of a Decision's Importance in Society

Facebook's curated newsfeed presents an alternative example where the role of the algorithm remained the same but the criticism came from acknowledging a change in the importance of the decision in society. Facebook came under scrutiny that their algorithmically curated Tending News was more liberal than conservative, and this bias was judged to be inappropriate. Facebook employees were intervening and removing articles that were poorly sourced or deemed unreliable. In response, Facebook removed the employees working on the trending topics section who were previously told to independently verify stories; the algorithm was delegated the task of deciding which stories were news and trending. In effect, Facebook increased the role of the algorithm in the creation of a newsfeed.[34]

However, Pew Research notes that over 50% of all Americans receive their news from Facebook, and for some in particularly restricted countries, Facebook is their only source of news.[35] While Facebook believed curating news had a minor role within society, society increasingly relied upon Facebook as an important news source. Facebook's solution—to remove the employees working on trending topics and rely

31 This example is highlighted in both O'Neil, C. op. cit., 2016, and Pasquale, F. "The Black Box Society: The Secret Algorithms That Control Money and Information," *Harvard University Press*, 2015.

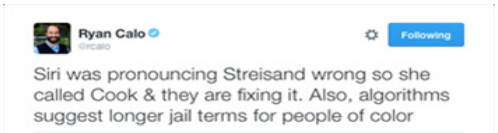
32 Burrell, J. op. cit., 2016.

33 As, I have noted elsewhere, the use of "it's complicated" by corporations has a long history of hiding malfeasance, including Enron and fracking, as well as credit default swaps and mortgage-backed securities. As Burrell notes, "Though a machine learning algorithm can be implemented simply in such a way that its logic is almost fully comprehensible, in practice, such an instance is unlikely to be particularly useful. Machine learning models that prove useful (specifically, in terms of the 'accuracy' of classification) possess a degree of unavoidable complexity," In my view, the instances of justifiable inscrutability are rare; moreover, designed-in inscrutability renders the developer responsible for algorithmic mistakes since individuals cannot identify, assess or correct mistakes in use.

34 See: (1) Tufekci, Z. op. cit., 2016; and (2) Dewey, C. "Facebook Has Repeatedly Trended Fake News since Firing Its Human Editors," *The Washington Post*, October 12, 2016, available at https://www.washingtonpost.com/news/the-intersect/wp/2016/10/12/facebook-has-repeatedly-trended-fake-news-since-firing-its-human-editors/?tid=sm_tw&utm_term=.b795225d264d.

35 See: (1) Gottfried, J. and Shearer, E. "News Use Across Social Media Platforms 2016," PewResearchCenter, May 26, 2016, available at http://assets.pewresearch.org/wp-content/uploads/sites/13/2016/05/PJ_2016.05.26_social-media-and-news_FINAL-1.pdf; and (2) Tufekci, Z. "The Real Bias Built in at Facebook," *The New York Times*, May 19, 2016, available at http://www.nytimes.com/2016/05/19/opinion/the-real-bias-built-in-at-facebook.html.

Figure 6: Tweet from Professor Ryan Calo, University of Washington, Law School)



solely on the algorithm—would appear to be the opposite of what we would find appropriate given the arguments here. Facebook may not want "editorial judgment over the content that's in your feed,"[36] but the role of its platform in providing a prioritized and validated news source may render that desire unimportant.

While autonomous AI may be possible, such an algorithmic decision may not be desirable for a particular decision context. Professor Ryan Calo, at the University of Washington School of Law with a focus on robotics, perhaps summarizes the dilemma best in the tweet shown in Figure 6: while Apple was willing to intervene in its artificial intelligence agent, Siri, to ensure Barbra Streisand's name was pronounced correctly, we are reluctant to have individuals intervene to take appropriate action in sentencing algorithms, even in the face of unjust biases.

Importantly, the design decision is possible; and this paper has examined the obligation of companies to actively engage in the ethics and accountability of algorithms in design. However, more work is necessary to understand the appropriate delegation of roles and responsibilities between algorithms and individuals, and under what circumstances.

Responsibility for Mistakes in Algorithmic Decisions

Algorithms create meaningful order out of large ambiguous datasets by sorting and prioritizing individuals. While previous attempts to categorize algorithms have focused on technical specifications or the type of output, here I suggest understanding algorithms based on the role of the algorithm in the decision and the importance of that decision in society. The development of algorithms is morally relevant in terms of not only creating mistakes but also in delegating the tasks of who can and should identify, judge, and correct mistakes in algorithmic decisions. In other words, algorithms are designed with a particular type of governance in mind. In effect, computer scientists design the role the algorithm has in the decision-making process and how much governance is possible by individuals. *If individuals are not given the opportunity to identify, judge, and correct mistakes as part of governing the algorithm in the decision, then developers preclude individuals from taking responsibility.* Table 2 offers questions for IT executives and algorithm development teams to navigate how ethical algorithms should be designed for use, based on the need to govern mistakes in algorithmic decisions.

I next ground why developers should take responsibility for designing the role of the algorithm in the decision: based on (1) the unique position of the computer scientist developing the algorithm, (2) the social contract entered into when the developer decides to become a member of the decision context, and (3) the identity of being a computer scientist.

Unique Position Argument

Development teams who design and develop algorithms, and therefore inscribe the preferred outcomes—including who can identify, judge and correct mistakes—create a system of decision-making that is difficult to undo or change in use. In fact, some firms intentionally create algorithms that are difficult to know and understand to gain a competitive advantage.[37] Other firms may use techniques that make it difficult to identify or understand the role of the algorithm, and thus make the algorithmic decision-making process

36 Solon, O. "Facebook Staff Mount Secret Push to Tackle Fake News, Reports Say," *The Guardian*, November 15, 2016, available at https://www.theguardian.com/technology/2016/nov/14/facebook-fake-news-us-election-news-feed-algorithm.

37 Larson, J. Mattu, S., Kirchner, L. and Angwin, J. How We Analyzed the COMPAS Recidivism Algorithm, *ProPublica,* May 23, 2016, available at https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Table 2: Questions for IT Executives and Development Teams to Ensure Ethical Algorithms

Questions for CIOs and CDOs	Questions for Development Teams
… about Identifying Mistakes:	
• What are the types of mistakes possible within the decision? How are the decisions going to be identified in use?	• What is the distribution and frequency of these types of mistakes in the algorithmic decision and in the (current) human-centric decision? Is the algorithmic decision better than the human-centric decision?
… about Judging Mistakes:	
• What are the risk tolerance and error ranges for mistakes, and the preference for a type of mistake based on the decision context? What type of mistake is preferred—if any?	• Are mistakes disproportionately falling on one group? Is this fair? • Are the mistakes appropriate within the norms or rules of the decision context?
…about Correcting Mistakes	
• What are the norms of the decision context about who should be correcting mistakes for the organization?	• How are mistakes fixed in a way that ensures future mistakes are not made? • How are mistakes corrected before the decision is implemented?
…about the Role and Accountability of the Algorithm in Decisions	
• Is the decision context pivotal? • How opinionated do you want your algorithm to be for this decision? How responsible do you want the developers and users to be for the outcomes of the algorithm?	• Is the level of automation appropriate for the decision context? What type of social embeddedness and reflection is necessary in the algorithmic decision? • Does the decision have strong norms in society?

inscrutable.[38] The unique knowledge or unique position argument is akin to the obligations of doctors to render aid—if programmers do not design algorithms to take into account how mistakes will be identified, judged, and corrected, then no one else will be able to. Developers are uniquely situated—with knowledge and position—to effect change in how algorithms can be governed. In creating the algorithm, developers are taking a stand on ethical issues and "expressing a view on how things ought to be or not to be, or what is good or bad, or desirable or undesirable."[39]

Members of Decision Context Argument

Similar to engineers needing to understand the best manufacturing practices when designing robots for manufacturing, algorithm developers also need to understand the norms of the algorithm-in-use as well as the best practices of ethical decision-making. In making the decision to sell an algorithm in a decision context to universities to sort applicants, to firms to sort job applicants, to courts to categorize defendant risks, etc., developers willingly enter into that community as a member of the decision system. And as a member of the community, that firm now has an obligation to understand the norms of the decision and not violate those norms in the use of the algorithm. If a company does not wish to make their algorithm understandable to the larger community, as was requested by defendants subject to the risk assessment algorithm used in sentencing, then the firm should not sell communities where due process is the norm. This is a social contract argument where the firm developing and selling the algorithm (and the actual computer scientists as members of that organization) take on the obligation of being good members of the community they willingly enter.

38 See: (1) Burrell, J. op. cit., 2016; and (2) Desai, R. D. and Kroll, J. A. "Trust but Verify: A Guide to Algorithms and the Law," *Harvard Journal of Law & Technology* (31:1), Spring 2018.

39 Kraemer, F., van Overveld, K. and Peterson, M. "Is There an Ethics of Algorithms?," *Ethics and Information Technology* (13:3), September 2011, pp. 251-260.

Identity as a Computer Scientist Argument

Finally, as computer scientists, and as engineers more broadly, developers of algorithms make value judgments their jobs as computer scientists and engineers. Philosopher Richard Rudner famously noted that scientists in their jobs as scientists make value judgments, and the job of the scientist includes proactively acknowledging and managing those value-laden decisions, such as which problems to solve, what is important to consider, what is a good result, etc.[40] Here, the argument is to similarly broaden what it means to be a good algorithm developer or computer scientist. *The job of a developer includes designing how the algorithm can and should be governed by individuals while in use.* Unattended mistakes are unethical and it is the obligation of developers-as-developers to ensure mistakes are governed. The very job of developing a good algorithm and the criteria of judging a good algorithm need to be broadened to include governance questions, such as design decisions about identifying, judging, and correcting mistakes.

Concluding Comments

Individuals and firms who develop algorithms make morally important decisions that are embedded in the algorithm with implications about who is accountable for identifying, judging, and correcting mistakes. Design decisions can cause an algorithm's role in a decision to be inflated, particularly if the algorithm is hidden, inscrutable, or autonomous. Ethical decision-making scholarship offers both social embeddedness and reflection as important attributes of good decisions and possible levers to deflate the enlarged role of algorithms in decision-making. All algorithmic decisions will produce mistakes; ethical algorithms will offer a mechanism to identify, judge, and correct mistakes. In this paper, I have argued that this design—of being comprehensible in terms of identifying, judging, and correcting mistakes—is indeed a decision and one for which developers of algorithms should be held accountable. By creating inscrutable algorithms, which are difficult to govern, developers may voluntarily take on accountability for the role of the algorithm in the decision.

40 Rudner, R. "The Scientist Qua Scientist Makes Value Judgments," *Philosophy of Science* (20:1), January 1953, pp. 1-6.

About the Authors

Kirsten Martin

Kirsten Martin is an associate professor of strategic management and public policy at the George Washington University's School of Business. She writes about privacy and the ethics of technology in leading academic journals across disciplines. Kirsten is the Technology and Business Ethics editor for the J*ournal of Business Ethics* and the recipient of three NSF grants for her work on privacy, technology, and ethics. She regularly speaks on privacy and the ethics of big data, including her recent Tedx talk. She earned her B.S. in Engineering from the University of Michigan and her M.B.A. and Ph.D. from the University of Virginia.