Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation

Article *in* MIS Quarterly · September 2021

DOI: 10.25300/MISQ/2021/16535

MIS Quarterly

SPECIAL ISSUE: MANAGING AI

FAILURES OF FAIRNESS IN AUTOMATION REQUIRE A DEEPER UNDERSTANDING OF HUMAN–ML AUGMENTATION[1]

Mike H. M. Teodorescu
Carroll School of Management, Boston College, Fulton 460, 140 Commonwealth Avenue,
Chestnut Hill, MA 02467 U.S.A. {mike.teodorescu@bc.edu}

Lily Morse
John Chambers College of Business and Economics, West Virginia University,
Morgantown, WV 26506 U.S.A. {lily.morse@mail.wvu.edu}

Yazeed Awwad
Center for Complex Systems, King Abdulaziz City for Science & Technology, Riyadh 12354 SAUDI ARABIA,
and Massachusetts Institute of Technology, 77 Massachusetts Avenue, Building E18-309,
Cambridge, MA 02139 U.S.A. {awwad@mit.edu}

Gerald C. Kane
Carroll School of Management, Boston College, Fulton 460, 140 Commonwealth Avenue,
Chestnut Hill, MA 02467 U.S.A. {gerald.kane@bc.edu}

Machine learning (ML) tools reduce the costs of performing repetitive, time-consuming tasks yet run the risk of introducing systematic unfairness into organizational processes. Automated approaches to achieving fairness often fail in complex situations, leading some researchers to suggest that human augmentation of ML tools is necessary. However, our current understanding of human–ML augmentation remains limited. In this paper, we argue that the Information Systems (IS) discipline needs a more sophisticated view of and research into human–ML augmentation. We introduce a typology of augmentation for fairness consisting of four quadrants: reactive oversight, proactive oversight, informed reliance, and supervised reliance. We identify significant intersections with previous IS research and distinct managerial approaches to fairness for each quadrant. Several potential research questions emerge from fundamental differences between ML tools trained on data and traditional IS built with code. IS researchers may discover that the differences of ML tools undermine some of the fundamental assumptions upon which classic IS theories and concepts rest. ML may require massive rethinking of significant portions of the corpus of IS research in light of these differences, representing an exciting frontier for research into human–ML augmentation in the years ahead that IS researchers should embrace.

Keywords: Fairness, machine learning, augmentation, automation, artificial intelligence

[1]Nicholas Berente, Bin Gu, Jan Recker, and Radhika Santhanam were the accepting senior editors for this paper. Michael Barrett served as the associate editor.

Mike Teodorescu and Lily Morse share first authorship for this paper.

Introduction

Since machine learning (ML) systems became widely available, organizations have considered the prospect of using ML models to increase productivity (Aghion et al. 2017). Yet when these models are applied uncritically, they can result in outcomes that unfairly advantage some groups of people while disadvantaging others. For example, researchers found that Facebook's ad targeting algorithms perpetuated gender disparities in job postings such that female users were less likely to see ads from companies that predominately hire male employees (Imana et al. 2021). When applying ML models in ways that may influence outcomes of socioeconomic importance, organizations need to invest care in ensuring that the automated technology is enacted *fairly* and does not discriminate. Fairness refers to treating others as one wishes to be treated, following agreed-upon societal standards (Ambrose and Schminke 2009; Cropanzano et al. 2003). Algorithmic fairness is a crucial issue in ML.

Yet, designing ML systems to achieve fairness automatically can be difficult in all but the most basic situations, and in some cases impossible. For example, Hao and Stray (2019) explore the complexities of achieving fairness in the COMPAS system, a system the U.S. justice system uses to determine the risk of recidivism. Different definitions of fairness that system designers could use are often mutually exclusive, result in different outcomes, and developers cannot resolve these differences. Yet, these differences play a critical role in deciding which prisoners are released and which prisoners stay jailed. Researchers have also identified numerous computational limits for developing fair algorithms in other settings, demonstrating that fairness is computationally intractable (Dinov 2016; ur Rehman et al. 2016). The computer science literature has developed dozens of fairness metrics, many of which are mutually incompatible, with little guidance on which developers should use and when (Chouldechova 2017; Mehrabi et al. 2019). Hao and Stray conclude their analysis of fairness in the COMPAS system, noting that "no algorithm can fix this; this isn't even an algorithmic problem."

On the other hand, simply removing ML from fairness decisions is not a solution for achieving fairness because humans exhibit considerable bias in their decision-making (Bazerman and Tenbrunsel 2012; Frank et al. 2019). Prior work has highlighted the various ways that people make counterintuitive and biased judgments in the workplace, including recruitment and hiring decisions (Rivera 2012), performance appraisals (Levy and Williams 2004), and financial assessments (Milkman et al. 2008). In other words, research suggests that neither humans nor ML models are likely to achieve fairness working alone. Instead, human–ML augmentation, where humans and technology work together to perform organizational tasks jointly, is the most promising path to achieving fairness.

Much of the existing literature advocating augmentation simply forwards the concept without exploring the practical details of how to achieve it, overlooking important situational and interactional information (e.g., Lindebaum et al. 2020; Raisch and Karkowski 2020). Attempts to provide a more detailed view of augmentation typically treat it as a function of the augmented technology rather than a function of the entire sociotechnical environment in which this augmentation occurs (e.g., Murray et al. 2021). The IS discipline has long recognized that the influence of technology on organizations is not merely a function of technology (e.g., Bostrom and Heinen 1997; Orlikowski and Scott 2008) but of many sociotechnical factors that may interact in complex (e.g., Nan 2011) and often unintended ways (e.g., Watson et al. 1988). Yet, little existing IS research explores these sociotechnical complexities of human–ML augmentation for fairness.

In this paper, we argue that human–ML augmentation is more complex and nuanced than is currently represented in the literature and more research is needed to understand these nuances better. To illustrate this need, we develop a typology of augmentation composed of two dimensions: the difficulty of achieving fairness on a given set of variables and the locus of decision in the human–ML partnership. This typology results in four distinct approaches to augmentation to achieve fairness: reactive oversight, proactive oversight, informed reliance, and supervised reliance. We first provide a definition and example for each form of augmentation. We then connect each augmentation type to a referent stream of earlier IS research, demonstrating that the IS literature has long pursued a nuanced view of human–technology augmentation and introduce specific research questions that arise within each category. We conclude by suggesting managerial strategies for achieving fairness in each of the four types of augmentation. Without such a subtle understanding of the fairness landscape, organizations risk applying the wrong fairness strategy in a particular situation, which could spiral into increasingly unfair treatment.

While we focus exclusively on augmentation for fairness throughout the paper, we hope this work will inspire other IS researchers to explore similar nuanced views of augmentation in different contexts, potentially opening a robust stream of research.

Background: Fairness Criteria

Early ML models aimed to recognize patterns and correlations in data without requiring knowledge of the nature of the phenomenon. The trend in ML is to establish quantitative models in applications that previously benefitted from qualitative theoretical models alone. ML helps to understand phenomena when the underlying theoretical model may be unknown and

when the phenomenon can be derived from the data and simulated to predict future events from past observations. This term's "learning" component represents the criterion that the algorithm outcome must improve with more data. Developers assess this improvement by performance criteria, under which the parameters of the algorithm must optimize (for an overview, see Mitchell 1997).

A key concept in algorithm fairness is *protected attributes*, which refers to individual characteristics protected by law and cannot be used to discriminate, algorithmically or otherwise: race, religion, gender, national origin, marital status, age, and socioeconomic status. A variety of laws define these protected attributes, including in U.S. laws on housing (the Fair Housing Act), credit (The U.S. Equal Credit Opportunity Act), disability rights (Americans with Disabilities Act), and hiring (Federal Equal Employment Opportunity Act).

In addition to concerns for societal well-being, these legal protections obligate firms to construct and implement fair algorithms that comply with defined legal statutes. Furthermore, unfair ML models can result in substantial fines for firms as well as reputational damage. Examples of documented unfair ML include Facebook's recent violation of the Fair Housing Act through ad targeting (Benner et al. 2019), Amazon's recruiting tool that discriminated against female coders (Dastin 2018; Meyer 2018), and Northpointe's COMPAS system, which discriminated by race and gender (Hao 2019). Table 1 describes four well-known fairness criteria from the computer science literature, emphasizing how these criteria may lead to model discrimination through their treatment of protected attributes. More elaborate criteria may require richer and more diverse training sets.

These multiple definitions of fairness pose challenges for automated approaches to fairness. Just as fairness does not have a universal definition among humans, fairness does not have an agreed-upon definition in computer science. Many of the fairness criteria from the computer science literature are mutually exclusive, in that several criteria cannot be satisfied at the same time (Mehrabi et al. 2019). Choosing the correct definition of fairness for a particular situation is challenging, and choosing multiple definitions may be impossible.

Failures of Fairness Through Automation

Automating fairness becomes even more challenging when seeking to optimize on more than one definition of fairness or more than one fairness attribute. The difficulties do not arise only from the nature of the problem landscape (e.g., Kauffman and Weinberger 1989) or from "the combined challenge of unforeseeable uncertainty … and high complexity" (Sommer and Loch 2004, p. 1334). Instead, it is a combination of these and the current limits of automation. Current ML systems do not have a good ability to deal with the "curse of dimensionality" (Altman and Krzywinski 2018; Dinov 2016). Machines may also be confused by wild outliers. For example, it is possible to fool facial recognition systems with just a few pixel changes (Hao and O'Neill 2020). Further, as combinations of fairness criteria are, under certain circumstances, mathematically impossible—the "impossibility theorem" (Chouldechova 2017; Kleinberg et al. 2016; Saravanakumar 2021)—this negates the hope that a weighted combination of them will do better.

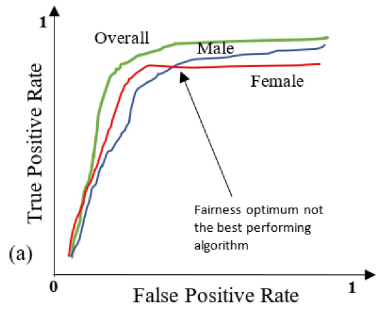
Even in the most straightforward case of a protected attribute with two categories (e.g., gender), there is a tradeoff between accuracy, an oft-used metric for model performance, and fairness. Figure 1 illustrates the receiver operating characteristic (ROC) curves for two genders, showing that the model minimizes errors for males and for females at different peaks. The ROC is a true positive rate versus false positive rate plot, which is a typical representation for classifier performance. The intersection between these curves would be the fair choice, although it achieves a lower accuracy than optimizing for one group versus another. The model may need to sacrifice accuracy for the sake of a fair outcome. Additionally, there may be no fairness optimum if the two ROC curves do not intersect. Figure 1b shows this outcome, where there is no group-level fair optimum, as the model performs so much worse on one gender versus another that there is no intercept between the two ROC curves.

The situation becomes even trickier with three or more subgroups. Can we achieve this intersection for every race in the sample? There is no fair optimum in many cases, as there would be no intercept for all the ROC curves delineated by each subgroup. Thus, human judgment regarding which criterion to apply and intuition from understanding the stakeholders and the distribution of protected attributes in the data is necessary to inform the choice of a fairness optimum. This level of human involvement is absent in pure automation. In practice, the developer or operator of the model, such as the hiring manager at a firm, should determine a non-optimum degree of fairness based on its specific goals. For example, a firm with historically low hiring of female software engineers may choose an optimum with a higher rate of false positives on female applicants' ROC curve in Figure 1b to increase diversity. The organization may adjust this optimum over time.

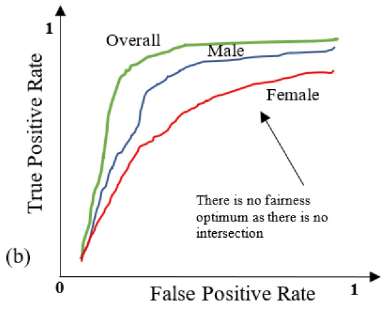
While group fairness is relatively intuitive to conceptualize (e.g., do the outcomes differ across a protected attribute?), its implementation is not. For instance, the satisfaction of a criterion for one protected attribute may be mutually exclusive

Table 1. Summary of Typical Fairness Criteria

Fairness Criterion	Definition	Limitations	Utility
Fairness through Unawareness	Fairness through unawareness involves leaving out protected attributes from the dataset (Kusner et al. 2017), and it is often the default fairness method (Chen et al. 2019; Kusner et al. 2017).	Possible correlations with protected attributes can exacerbate existing biases while giving the impression that the machine has acted fairly.	It is very limited: it would only apply in the unlikely scenario of no correlation between protected attributes and the features used to predict outcomes.
Demographic Parity	Equivalent to independence of the outcome for the protected attribute: $p\left(\hat{Y} \mid A = a\right) = p\left(\hat{Y} \mid A = a'\right), \hat{Y} \perp A$ where $\hat{Y} \perp A$ denotes independence, and a and a' are any couple of values of the protected attribute A and $\hat{Y}$ denotes the predicted outcome (Kusner et al. 2017).	Demographic parity can compound bias against those who are members of multiple protected groups.	If interested in achieving just parity for only one protected attribute, demographic parity is a better approach than unawareness.
Equalized Odds	Requires that for every value of the protected attribute, both the true positive rate and the false positive rate are the same (Kusner et al. 2017): $p\left(\hat{Y} \mid A = 0, Y = y\right) = p\left(\hat{Y} \mid A = 1, Y = y\right)$	Each protected attribute requires an additional test of the criterion, which can be challenging if we have multiple protected attributes in the data (strict equality may be impossible).	Equalized odds enforces equality of error rates across the protected attribute, not just the same outcome, providing a stronger group fairness metric than prior two (Hardt et al. 2016; Kilbertus et al. 2017). However, as a stricter criterion, it has more limited utility.
Equalized Opportunity	The equalized opportunity principle states that positive outcomes should be independent of the protected attribute (Hardt et al. 2016; Kusner et al. 2017). $p\left(\hat{Y} = 1 \mid A = 0, Y = 1\right) = p\left(\hat{Y} = 1 \mid A = 1, Y = 1\right)$	Equalized opportunity is a restricted case of equalized odds and is more easily achievable.	As a weaker version (assumption-wise) of equalized odds, it allows for stronger utility over more contexts (Hardt et al. 2016) than equalized odds, providing the best balance of fairness and utility.



(a) Example in which, for two groups, we can achieve a fairness optimum (group-level), but that is not the best-performing model (adapted from Awwad et al. 2020).



(b) Example in which there is no fairness optimum, as the performance of the classifier is much worse for female applicants than for male applicants.

Figure 1. Examples for Fairness Optimum

with the satisfaction of a criterion for another, requiring the developer to choose which attribute to improve upon, as there is often no "globally fair" solution. The impossibility theorem states that if multiple groups differ in their outcomes, certain types of fairness are unachievable. For example, it is impossible to equalize across any three metrics simultaneously because every three becomes mutually exclusive, including demographic parity, false-positive rate, and false-negative rate (Chouldechova 2017). Even for the most straightforward criteria of matching group-level outcomes, a developer must choose which criteria to equalize. Picking some criteria will mean that satisfying others will be impossible.

Further, in the case where group fairness has been achieved (on whichever metric the implementer of the system selects), a composition of group-fair classifiers does not yield a group-fair system post-composition (Hardt et al. 2016). In other words, combining two separate fair classifiers on different data into one does not result in fair outcomes. Some of the implementations used in practice are also known to create strong unfairness at the group level by including multiple criteria for fairness without a unique solution across them (e.g., "fairness gerrymandering"; Kearns et al. 2018). Yet, when dealing with more than one protected group, such as gender and race, auditing fairness *ex post* can become computationally infeasible (Kearns et al. 2018).

Simply put, fairness criteria are often incompatible. In Figure 2, we illustrate mutually exclusive fairness criteria through several visual examples. The scenario in Figure 2a shows four subgroups with their identifiable cluster by protected attributes. The clustering is present in the data before running any classification task, as exemplified in the rightmost panel. Since this visualization approach makes it easier to reveal the flaws in a solution that divides by attribute between, rather than across, the distinct clusters, simply visualizing the classification boundary and showing the protected attributes' values would indicate whether group-level fairness across race and gender is respected. For simplicity, we illustrate two categories for each attribute.

On the leftmost panels of Figures 2a and 2b, the data show that Black males and females are essentially a single group. In such a case, group-based fairness will only treat visible clusters with parity. Under group-level fairness, such a blending of the categories is not possible (comparing the three categories Black, White males, and White females instead of gender and race, which forces four categories). The two "cuts" in Figure 2b illustrate the group-fairness enforcement approach, which suits the given data well, but would fail against the previous data from Figure 2a in which there was a single cluster for Black males and females. For group-level fairness, the choice often boils down to a crudely binary conception of fairness that looks at one attribute at a time versus a burdensome pairwise bundling of attributes and criteria that grows exponentially with the number of categories, whether those pairings reveal meaningful relationships or not.

In summary, more complex fairness decisions under automation present considerable challenges. The common perception that designers can achieve fairness simply by incorporating fairness criteria with proper weights into the objective function will not achieve fairness in all but the most basic settings because combinations of fairness criteria are mutually exclusive. Further, increased complexity can lead to failure of the algorithm, hence the danger of a hands-off approach to managing ML models. If automated methods alone are unlikely to ensure fairness, then successful approaches are likely to involve augmentation.

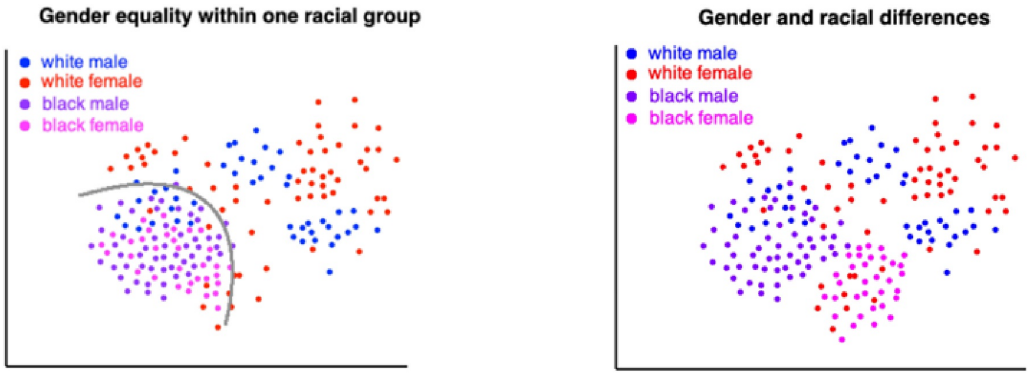


Figure 2a. Differences in Data across Gender and Race (The horizontal and vertical axes represent predictors, whereas the legend labels protected attribute combinations.)

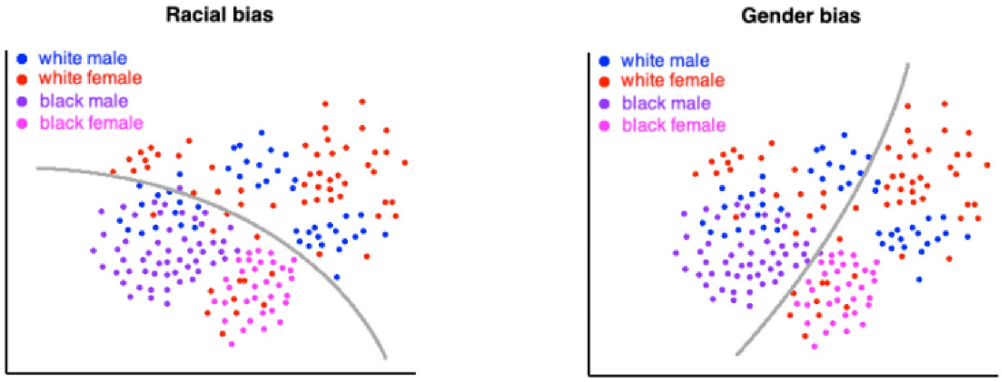


Figure 2b. Illustrations of Bias Based on Race (left panel) and Gender (right panel)

Figure 2. An Illustration of Mutually Exclusive Fairness Criteria

Realizing Fairness Through Augmentation

To shed light on the realization of fairness through augmentation, we consider recent organizational scholarship that has put forward strong arguments in favor of this approach over automation (e.g., Daugherty and Wilson 2017; Lindebaum et al. 2020). Augmentation is successful because it allows humans and ML to complement each other by relying on their strengths and overcoming their weaknesses. For example, humans can help offset many of the limitations of ML we identified at the beginning of the previous section. Humans can overcome the "impossibility theorem" by picking a solution that can be acceptably fair on a case-by-case basis and derive this solution from an understanding of the societal roots of unfairness (Cooper and Abrams 2021). Humans can help overcome the dimensionality problem (e.g., Kauffman and Weinberger 1989; Sommer and Loch 2004) by helping simplify the problem space by pinpointing salient aspects and removing futile others. Humans may also be able to identify wild outliers more easily than ML, not as easily fooled by imperceptible changes.

Other scholars have suggested that humans and ML excel in different areas of decision-making (Agrawal et al. 2017). ML models are superior to humans in *prediction*, which is particularly valuable in situations with a high degree of complexity because of their ability to analyze vast amounts of data. In contrast, humans are superior to ML in *judgment*, understanding the impact of the different outcomes generated by the prediction and making choices regarding that impact. Achieving fairness involves prediction and judgment: prediction about the likelihood of certain outcomes and judgment to manage the tradeoffs in situations that cannot operationalize with a single, measurable variable.

Augmentation can also help identify the social context, which leads to label bias in training data that developers cannot resolve with purely algorithmic approaches to fairness (Cooper and Abrams 2021). Without human intuition and under-

standing of what led to the bias in the "ground truth" labels (Wick et al. 2020), ML could end up mimicking an unfair set of past patterns. Augmentation brings the supplementary expertise of the human, those collecting the training set, and the programmers of the ML to bear on the problem.

Toward a Deeper Understanding of Augmentation for Fairness

This emergent literature conceptualizes augmentation as a collaborative partnership in which machines enhance human performance rather than replace it (Raisch and Karkowski 2020). Yet this simple definition leaves little room for the existence of complexity in the process of achieving fairness. Because the nature of human–ML partnerships is multifaceted, with multiple possible ways of shaping collaborative outcomes across decision-makers, researchers need to develop an evolved understanding of augmentation. We propose that managing fairness through augmentation is a complex process requiring different approaches depending on the context.

Over the past 50 years, the IS literature has produced significant insights into the complex and, at times, contradictory interactions between technology and organizations (e.g., Bostrom and Heinen 1977; Kane and Alavi 2008; Leonardi 2011; Orlikowski and Scott 2008). Against this background, we advocate a finer-grained understanding of augmentation for achieving fairness, one that accommodates a broader set of interactions between tasks, humans, and systems. When organizations overlook situations where there are complicated or uncertain interactions within the human–ML partnership, they risk developing a narrow mindset about augmentation, leading to poorly executed strategies for achieving fairness or other outcomes. In Figure 3, we introduce a typology of augmentation for fairness in which the nature of the human–ML partnership depends upon two underlying factors: fairness difficulty and locus of decision.

As displayed on the vertical axis, the first dimension is fairness difficulty, representing the number of fairness criteria the organization seeks to optimize. As the number of fairness metrics applied increases, so does the difficulty of achieving fairness in that setting. More attributes increase the fairness difficulty by increasing both the importance of prediction (i.e., it involves more data than humans can reliably handle) and judgment (i.e., it increases the need for tradeoffs between parameters that ML cannot make) in the fairness equation, making the role of human–ML augmentation more critical. Situations of high fairness difficulty overwhelm the strategies for augmentation advocated in low difficulty settings, requiring fundamentally different strategies for achieving fairness. As displayed on the horizontal axis, the second dimension captures the locus of decision, which describes who the final decision maker is in a particular situation, the human or the ML model. Unlike our fairness difficulty variable, a gradient based on the number of criteria, our fairness decision category is bimodal because only one partner can dictate the final decision.

These axes give rise to four quadrants that depict how organizations can approach fairness through augmentation: reactive oversight, proactive oversight, informed reliance, and supervised reliance. These categories do not represent mutually exclusive forms of augmentation; instead, they are broad conceptual groupings. Further, these approaches are robust in that researchers and managers can apply them to various situations. Each form of augmentation has a respective referent in the established IS literature. Yet, we believe that each category raises new questions about how augmentation should be applied in the context of ML, suggesting avenues for future research. By mapping augmentation using our matrix, organizations can more precisely identify the nature of their partnership with ML tools. They can better design for accountability in complex technological environments, in which it is often challenging to achieve fairness.

ML as the Locus of Decision

First, we will explore the left column of the matrix, where the ML model is the final decision-maker, and augmentation mainly involves human oversight of the model's decision. In this situation, people monitor the model's decisions, providing supervision to ensure that it does not begin to act unfairly. The nature of the oversight that individuals offer may differ in situations of low versus high fairness difficulty. Under conditions of low fairness difficulty, people may simply monitor and supervise the model's decisions, reacting by stepping in to adjust or retrain the model if they discover that decisions have become unfair. However, when fairness difficulty is high, it may be more challenging to recognize where unfairness is occurring and retrain the model to compensate for it, creating a need to proactively guide the model's decisions to steer the machine toward fairer outcomes. We explore these types of augmentation in greater depth in the following two sections.

Reactive Oversight

The lower left quadrant of Figure 3 depicts augmentation involving low fairness difficulty in which the ML model is the final decision-maker. We refer to this type of augmentation as *reactive oversight*. We call it reactive because humans monitor the model's decisions *post hoc*, only seeking to

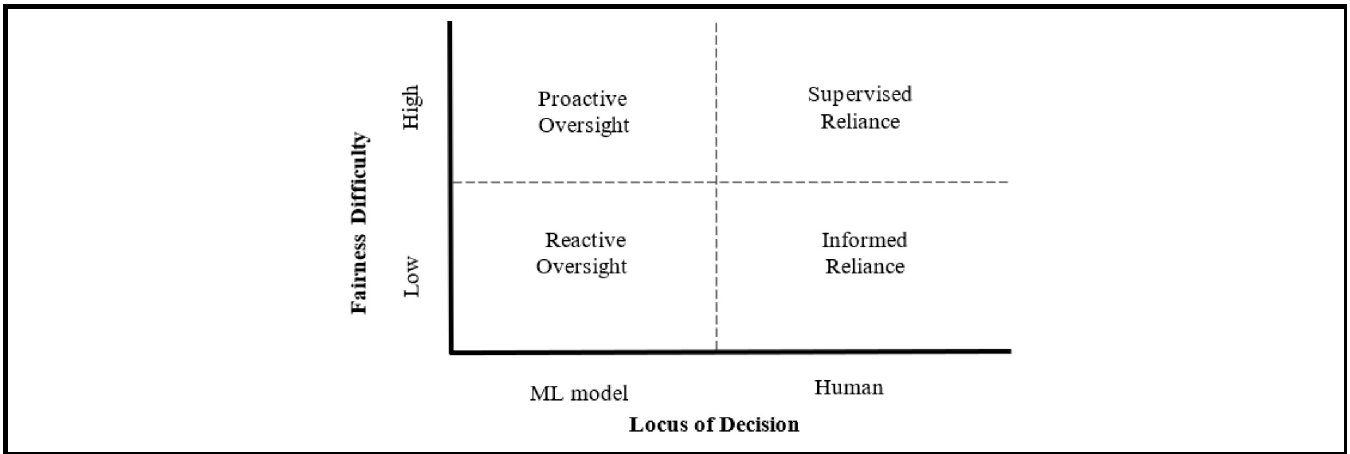


Figure 3. Typology of Augmentation for Fairness

modify decisions when the model begins to act in unfair ways. For instance, developers could improve the model's performance upon noticing unfairness by tweaking its decision parameters, such as the fairness/accuracy threshold. Reactive oversight focuses on employing existing fairness criteria and analyzing how well the ML model adheres to these criteria.

An illustration of reactive oversight is financial technology (fintech) companies, where they use ML models for identity verification in validating loan worthiness. Because the United States closely regulates lending bias, the government can substantially fine companies if these identity verification models contain biases in age, gender, or socioeconomic status that would introduce discrimination into lending decisions (Kane et al. 2019). Therefore, humans are continually auditing these models' outcomes to ensure that they are not acting unfairly and modifying them if the human monitors detect unfairness.

Significance for IS Research. Reactive oversight has some referents in the existing IS literature. For instance, security matching (e.g., August and Tunca 2006; Cavusoglu et al. 2008) is one way that humans supervise systems in operation and modify them as necessary. When software vendors realize that their products have security vulnerabilities, they typically release "patches" that allow customers to fix these vulnerabilities. Some research shows that releasing imperfect software early and applying these patches when developers identify imperfections during operation is a superior strategy compared to waiting to release versions without these imperfections (Arora et al. 2006). Likewise, companies can implement fixes and retrain ML models when fairness violations are identified rather than waiting for models to be sufficiently fair before relying on them. Borrowing from the security literature, we might refer to these procedures as "fairness patches." There is a need for future research to expand our understanding of reactive oversight. It is unclear when and under what conditions human monitors should react to address fairness concerns. For instance, it may be essential to consider time as a variable. Stepping in too late when a model's outcomes are unfair could inflict considerable damage on affected stakeholders. In contrast, intervening to restore fairness before there is enough knowledge of the underlying issue could lead to similarly disastrous outcomes creating more unfairness for underrepresented people. Reacting to unfairness is also considerably more challenging for ML tools trained on data than built from code (see Kane et al. 2021). Developers typically must completely retrain the model, rather than just adjusting the parts of the code that lead to unfair outcomes.

Managerial Strategies for Reactive Oversight. The managerial strategies for reactive oversight involve providing clear fairness objectives and incentives for ML development teams to achieve those objectives. Managers may want to specify in advance which definitions of fairness they want to adopt and which groups are more important to prioritize for fairness. Recognizing that perfect fairness is likely impossible to achieve, it may be preferable for managers to execute well on more modest fairness goals than to fail at accomplishing more comprehensive ones. For example, focused monitoring that provides intermittent performance feedback and clearly emphasizes the priority level of fairness over potentially conflicting goals, such as accuracy, may improve ML models' quality of human oversight and actual performance. Business ethics scholars have also called for more clarity in defining accountability for those in charge of algorithms, including being more transparent in specifying their level of involvement and oversight of the model (Martin 2019).

Proactive Oversight

The upper left quadrant of Figure 3 depicts a situation in which the ML model is the final decision-maker in high fairness difficulty situations. *Proactive oversight* occurs when the ML model makes final decisions on outcomes, and multiple fairness metrics are applied to achieve fairness. Because the fairness task is difficult—the model may be accountable for managing various interests underlying the fairness issue at hand (e.g., subgroup fairness)—the model must identify the best path out of numerous possible pathways for achieving an acceptable solution. Because it is challenging, if not impossible, to achieve perfectly fair outcomes in high difficulty situations (see Figures 1 and 2), the human actively intervenes to guide the ML model toward fairer decisions based on the tradeoffs faced in particular contexts. Thus, we refer to this quadrant as proactive because the human is not waiting to identify unfairness to act but is guiding the ML model toward fairness in an ongoing way. Further complicating the situation, interactions with decision-makers who attempt to steer the model toward a particular direction that aligns with their desired fairness outcomes at the expense of fairness for others may also influence the ML model. Consequently, it is incumbent upon organizations to monitor the ML model's decisions, react to unfairness when identified, and, most importantly, provide active stewardship and ongoing feedback to the system to help the model realize globally optimal fairness solutions.

Content filtering in online platforms is one example of proactive oversight. These sites rely on ML models to capture and integrate various user data to create personalized newsfeeds. Without active stewardship from developers or those in charge of the model, people who prefer to reinforce views similar to their own may attempt to provide model feedback in such a way that the ML model begins to move toward their version of fairness. This situation potentially creates ideological echo chambers that overtly discriminate against certain groups (Wolf et al. 2017). Some people may misuse the platform to spread false or discriminatory posts (Del Vicario et al. 2016). In this case, an ML moderator should decide which posts are unfair, such as posts that include inflammatory or discriminatory content against groups other than those promoting the post. Substantial public and scientific debates have taken place regarding the fair representation of information and how ML models can preserve multiple opinions on newsfeeds while avoiding misinformation (Frenkel et al. 2020; Shore et al. 2018). We add to this discussion by arguing for proactive oversight mechanisms built into ML models in this context. Because there is no straightforward decision formula for ensuring fairness in this type of augmentation, and because reacting to every instance of unfairness would be a Sisyphean task, the primary strategy should be to guide ML models toward fairer outcomes when possible. This strategy may involve manually overseeing certain types of feedback (e.g., liking or disliking a political ad) and guiding the models toward the pathways that come closest to meeting agreed-upon standards of right and wrong.

Significance for IS Research. The IS literature on workarounds serves as a reference for proactive oversight (Ferneley and Sobreperez 2006). In these situations, users seek to adapt a poorly designed system that may interfere with their ability to perform essential tasks. Consequently, people develop the workarounds—"informal temporary practices for handling exceptions to workflow" (Kobayashi et al. 2005, p. 1561)—that let them avoid problematic aspects of the system to accomplish necessary tasks or accomplish them more efficiently. We view proactive oversight as similar to this workaround behavior. Decision-makers enact temporary behaviors that override or ignore specific ML decisions when they perceive that the system could be acting more fairly in a given situation. The critical difference between proactive oversight and previous instances of resistance is that ML tools can learn from this constructive resistance and change their decision-making in response to feedback from the human overseers.

Further research is necessary to understand the implications of proactive oversight. Given that resistance is likely under this type of augmentation, a fruitful line of work would explore how fairness criteria and their implementation are likely to be perceived by different stakeholders in a given situation (e.g., developers versus consumers). Because people may try to transform or use particular ML model features to align them more closely with their ideals, it would also be interesting to examine *how* people guide the model and which tactics are most effective in promoting one's version of fairness. Finally, we discern the need to expand the conceptualization of fairness when it is narrow and to consider more global, balanced solutions for attaining complex outcomes.

Managerial Strategies for Proactive Oversight. The managerial strategies for proactive oversight provide support for nontechnical managers who need to be actively involved in this process. Not only would training and support be necessary to equip managers for these tasks, but it will also be necessary for managers to establish deterrent mechanisms for those who oversteer the model and misuse the system. For example, organizational research suggests using strong sanctions, such as punishment (e.g., banning people from the system), to prevent wrongdoing (D'Arcy et al. 2009; Tjosvold 1986). Strong sanctions have the most significant impact on situations where there is a "push" to behave selfishly (Smith-Crowe et al. 2015). Other research likewise recommends that organizations utilize coercive tactics to manage potentially exploitative employees in such situations by highlighting the negative consequences of failing to adhere to fairness goals (Tjosvold 1986).

Humans as the Locus of Decision

Next, we will explore the matrix's right column, where the human is the final decision-maker, and augmentation mainly involves deciding when to rely on an ML model's recommendation or guidance. Here, the fairness challenge is for people to be aware of the different ways that depending on the ML model's recommendation may lead to unfair outcomes and correspondingly adjust their final decision to account for this potential unfairness. It situates both the responsibility (i.e., accepting the potential costs, duties, and obligations of one's decision) and accountability (i.e., being answerable to others for decisions made and actions taken) for fairness with the human decision-maker (Recker 2022). When fairness difficulty is low, it may be sufficient to exhibit *informed reliance*. Decision-makers remain aware of how the model may provide unfair recommendations and use this knowledge to inform their final decision. However, when fairness difficulty is high, it may be impossible for individuals to remain aware of the myriad ways the model may provide unfair recommendations. In this situation, it will be necessary to provide additional support to the decision-maker to improve fairness through other algorithmic or human feedback mechanisms that we call *supervised reliance*.

Informed Reliance

The lower right quadrant of Figure 3 depicts augmentation in which the human is the final decision-maker, but there is lower fairness difficulty. *Informed reliance* occurs where individuals make the final judgment, and the fairness difficulty is relatively low. In this situation, decision-makers relying on the ML model for decision support simply need to be aware of the potential that the model will provide unfair recommendations and recognize that they must choose whether to rely on the model output when making decisions. We call this quadrant informed reliance because the humans' critical understanding of how those recommendations may be biased guides their decision to rely on the ML model's recommendations or not.

For example, a dermatologist may refer to convolutional neural networks (CNNs) before making a health diagnosis of skin lesions, including melanoma (Winkler et al. 2020) and skin infections (Fletcher et al. 2019). To engage in informed reliance, the doctor would need to be aware that the data on which the CNN was trained, like the data used to train many such ML tools, may have included relatively few minority patients (e.g., Awwad et al. 2020). It may be that only a limited number of skin tones are in the training set, leading the model to underperform on minorities. Thus, physicians may adjust their willingness to rely on the ML model's recommendation when treating minority patients, using their expertise to determine whether to accept the proposed solution or seek alternative recommendations.

Significance for IS Research. Prior IS research on expert or recommender systems is a referent for informed reliance in ML. Expert systems are computer programs that perform recommendations for specialized tasks based on understanding how human experts behave. Considerable research has investigated the factors associated with a user's willingness to rely on the system's recommendations (Meyer and Curley 1991; Xiao and Benbasat 2007; Ye and Johnson 1995). For example, rather than replace human decision-making, these recommendations function as assistive technical agents that are likely to be embraced by users when the reasons for the recommendation are straightforward and readily explained (Wang and Benbasat 2007).

Nevertheless, several outstanding research questions remain concerning this form of augmentation. The black-box nature of ML makes it challenging to explain why the model made the recommendation it did, potentially lowering cognitive and emotional trust in the system (Glikson and Woolley 2020). Understanding why decision-makers do or do not accept ML models' recommendations in their final decision will help develop informed reliance. In the fairness context, it may be just as important to understand the conditions that lead decision-makers not to rely on a recommendation as it is to understand the antecedents when they do. For instance, knowing why a dermatologist did not accept a CNN model's recommendation when treating a minority patient may be necessary for informed reliance because feedback to the algorithm designer on decision-maker rejection could improve the fairness of future iterations of the system. Furthermore, when individuals selectively choose which model recommendation to accept, they may demonstrate a preference for integrating decisions that favor personally held beliefs, expectations, or desired conclusions, leading to confirmation bias (Jonas et al. 2001). We identify the need for scholars to explore the potential problem of "cherry-picking" recommendations from assistive ML models more deeply.

Managerial Strategies for Informed Reliance. The managerial strategies for informed reliance involve educating and incentivizing human decision-makers to achieve fairness. Managers will need to provide learning opportunities to help facilitate decision-makers' technological literacy to better understand ML tools' general functioning and potential limitations (Kane et al. 2019). Only through understanding how the ML model may be systematically unfair can the decision-maker critically analyze those recommendations for potential unfairness. Some research similarly suggests that managers can use supportive tactics when managing employees to develop informed reliance (Gundlach and Cadotte 1994; Tjosvold 1986). Examples of supportive tactics in human–

ML interactions include offering advice to help employees carry out a fairness task, offering constructive feedback, and informally rewarding employees for displaying desired fairness-oriented behaviors (Yukl and Michel 2006). To facilitate trust-building between decision-makers and assistive ML tools, managers should consider providing nonmonetary incentives to employees, such as noticing and verbally appreciating those who successfully adopt fair recommendations by ML models (O'Donnell 2000).

Supervised Reliance

The upper right quadrant of Figure 3 depicts augmentation in which the human is the final decision-maker in situations of high fairness difficulty. We call it *supervised reliance* because, although the human is making the final decision, supervision of those decisions by algorithmic or human mechanisms is necessary to ensure fairness. Informed reliance will not suffice in this situation because people cannot remain aware of the myriad ways the ML model's recommendations may be unfair; thus, ongoing checks on the quality of humans' decisions are necessary. Much like individuals monitor ML models' decisions for unfairness in reactive oversight, here machines or people monitor human decision-making outcomes to help achieve fairness. While humans train ML models in supervised learning, other systems seek to train humans to make fairer decisions in supervised reliance. Supervised reliance differs from the types of oversight discussed in the previous section because it can only provide feedback and guidance to the human who ultimately determining future decision-making. In other words, a manager cannot fundamentally "tweak" a human decision-maker in the same way a development team can retrain an algorithm.

As an illustration, consider the HireVue platform, which tracks job candidates' behavior in video interviews (e.g., choices of words, facial expressions) to assess their "employability" and compares this data to a client company's current top performers. A criticism of this ML model is that it can easily discriminate against protected populations, such as older workers who may use different words in interviews or may be less comfortable interacting with interview technology (Barnes 2019). The model may recommend candidates because of characteristics such as agreeableness, conscientiousness, or vocabulary breadth. On the other hand, it might also make recommendations based on protected factors such as age, gender, or race. It will be difficult for human decision-makers to know which characteristics the ML model used to decide. Likewise, the model could simply optimize past decisions, which may be difficult for the decision-maker to recognize. For example, if a client company is predominantly composed of men, a manager might not realize that the model is more likely to recommend men for new positions. The HireVue example further reveals that although people might use ML models to attain fairer outcomes, there is a substantive risk that human error and bias will sway the model's recommendations, especially given the ambiguous nature of the environment. As the ML model learns which of its recommendations are accepted or rejected by the decision-maker, the model can reinforce these biases, potentially leading to a downward spiral of prejudice and inequity. In supervised reliance, machines or people would monitor the hiring manager's reliance on the HireVue algorithm to ensure he or she is not making unfair hiring decisions.

Significance for IS Research. Research on technology-in-use is a referent for supervised reliance (e.g., Orlikowski 2000). This perspective argues that the rules and resources embedded in information systems interact with the facilities, norms, and interpretive scheme embedded in individuals and organizations to develop a distinctive technology-in-practice. As a result, an information system embedded in one organizational environment is functionally different from the same system embedded in a different environment. These insights can help predict how ML models might evolve in unfair ways once in use, even if the models were fair when implemented initially. ML, however, takes this practice perspective of technology to another level. ML models function differently in different organizational contexts due to the organizational facilities, norms, and interpretive schemes. Still, the models can learn from these environments and modify their own embedded rules and resources. When the ML tools learn from the decision-makers' previous decisions, certain types of unfairness may be recognized and amplified by the models. Since the ML models may optimize on particular decision-makers' biases in a specific environment, it may be difficult for decision-makers to identify when these biases influence recommendations.

Future research should consider how both people and machines can effectively promote or correct human decision-making in supervised reliance. Perhaps developers can train a separate ML model to query a decision-maker when the current model's data is uninformative for fairness. Developers can also train it to identify when the current model targets wrongly predicted data points, such as active learning. Alternatively, researchers might explore how trained professionals can independently evaluate the quality of a human's decisions. The idea of humans in the loop may expect more from humans than they can readily deliver, just as the technical solutions for fairness address more than they can provide. Organizations can offer additional checks and fairness considerations beyond what the decision-maker had in mind to improve the overall outcome of fairness and firm performance. Additionally, research is needed to determine whether, how, and under what conditions human decision-makers use these checks to improve supervised reliance.

Managerial Strategies for Supervised Reliance. Managerial strategies for supervised reliance involve setting up processes to audit human decisions for potential unfairness and intervention techniques to increase fairness in those decisions. Managers can also select appropriate technological tools to monitor the human's decision-making and provide constructive feedback and solutions for how the human relies on the ML system to promote fairness. Given this human–ML partnership's complexity, managers should actively recognize and address decision and environmental uncertainty about fairness. For example, decision-makers should explicitly clarify what they know versus what they do not about the current situation, including the ML model, and determine whether their ambiguity for fairness is more attributable to a lack of awareness or understanding (Pich et al. 2002). Related scholarship in management and behavioral ethics offers various evidence-based strategies for reducing uncertainty in fairness contexts, such as designing organizational procedures that give clear fairness messages (van den Bos and Lind 2002).

Discussion: Generalizing Augmentation Beyond Fairness

In this paper, we explored fairness for the use of ML models in organizations. Despite recent attempts to develop fairer models, we show that fairness can rarely be fully automated. Instead, organizations might best achieve fairness through some sort of augmented human–ML partnership, which may balance with automation under the right circumstances. Previous papers that describe these augmented human–ML partnerships often oversimplify the associated complexities and difficulties that have had analogs in decades of IS research. We propose that a more robust treatment of these partnerships is necessary to leverage human–ML augmentation in organizational settings fully. As a first step, we develop a framework for understanding different types of augmentation based on the difficulty of fairness and the locus of decision. This framework results in four broad types of augmentation, each presenting unique problems, research questions, and management strategies for organizations and society.

Theoretical Implications

While we have focused on augmentation concerning fairness here, our insights can be generalized to provide insight into other forms of augmentation by loosening some of the assumptions specific to the fairness context. For example, we have defined difficulty in the fairness context as the number of fairness criteria the algorithm seeks to optimize. For other outcomes, this difficulty may take on additional forms. Certain outcome variables may be inherently abstract and not easily reducible to a single criterion or simple ethical desideratum or measurable outcomes—such as privacy (Lee et al. 2011), employability (Van Huynh et al. 2020), cultural fit (Lu et al. 2019), admissibility to college (Fong et al. 2009), or freedom (Kane et al. 2021). There may be only one outcome on which the ML model seeks to optimize, but that outcome may be difficult to capture no matter how many variables one uses to measure it. Outcome variables with differing degrees of complexity may appear at various points along the y-axis when determining the difficulty of optimizing other types of outcomes.

Additionally, we limited the discussion regarding the locus of decision to simply whether the human or the ML model makes the final decision in the fairness context. However, how the augmented partnership frames a problem or selects recommendations exerts a more significant influence on the outcome than the final decision. For example, the choice of data used to train the ML model may have a much more substantial impact on the outcome than the final decision, as might the number and range of options the model provides to a human decision-maker. Thus, as we generalize our framework, we might broaden the locus of decision axis beyond the final decision to the locus of influence, describing each partner's overall impact on the outcome. This change implies that this axis moves from a bimodal distinction representing who makes the final decision to a continuum describing each partner's influence over the entire decision process.

Our ability to connect each of our quadrants to various antecedents in earlier IS literature also has important implications for IS research. In several instances, we showed how differences between ML and traditional code-based systems either raise crucial new research questions or require us to revisit old streams of research in light of these differences. For example, how will decision-makers trust expert systems when those systems cannot explain the rationale for their recommendation? How do theories of technology-in-use change when the technologies themselves change depending on how people use them?

Furthermore, many of these differences between ML and traditional IS fundamentally undermine many assumptions of previous generations of IS theories (see Kane et al. 2014). For example, we noted that ML models generally need to be entirely retrained when unfairness is identified, not simply "tweaked" like we could with previous generations of code-based systems. What implication does this shift have on various theories and models for software development? Previous research also established that IS could only be valuable because individuals and organizations use them in particular settings (e.g., Goodhue and Thompson 1995). Yet, when the

ML system makes the final decision on fairness questions, these tools may add value or harm to the organization, even when no one explicitly "uses" them in a traditional sense. We have intentionally avoided the term "user" throughout this paper, where possible, to reflect this fundamental difference between how humans rely on ML differently than traditional IS. This distinction may open up an entirely new stream of research into how "effective use" might differ in the context of ML (Burton-Jones and Grange 2013).

These examples only scratch the surface of how IS researchers may need to reinterpret traditional theories and constructs in the face of ML. It represents a significant opportunity for IS researchers in coming years to open up new vistas of research into how these data-based ML tools shift our understanding of how IS operates in organizations. It also presents considerable risks if we do not employ established IS theories critically to understand these new tools, carefully considering whether and how researchers should revisit these theories in light of these changes.

Managerial Implications

In addition to the specific managerial implications mentioned in each subsection above, this paper provides two broad insights for practitioners. First, exclusively automated approaches to fairness are not likely to result in most organizations' desired outcomes. While we are not advocating that managers abandon automated approaches to fairness, they do need to recognize that such approaches are, at best, an incomplete solution. Instead, managers should increase their efforts at augmentation. Part of this augmentation effort will likely involve educating managers to understand the potential benefits and possible risks associated with ML. Only if managers adequately understand these issues can they begin to support the type of oversight and cultivate the kind of informed reliance necessary to achieve or improve augmented approaches to fairness. Furthermore, this understanding will help managers determine the right balance between performance outcomes and fairness outcomes in any given application of ML (see Figure 1a).

Additionally, because our framework is grounded in multiple characteristics of the human–ML relationship, it implies that humans may augment a single ML system in more than one way. For example, in our discussion of informed reliance, we described a dermatologist using their judgment when relying on a CNN's recommendations because they might be generally aware of potential bias in the system's recommendations. Nevertheless, the hospital's quality assurance manager may also use supervised reliance to ensure that the physician's informed reliance results in fair outcomes. The developers of that CNN might also engage in reactive oversight with that system to make the initial recommendations as unbiased as possible. The inference here is that no single type of augmentation is suitable for a particular system (see Murray et al. 2021): managers might apply many or all types to augment one system in various ways. Managers' strategy to support and guide augmentation will vary depending on the particular kind of augmentation they engage, not on the system itself. There is no "one size fits all" approach to managing fairness through augmentation for ML. Managers should be aware of these nuances as they choose the most effective strategy for a particular situation.

Conclusion

We believe that the IS discipline should move aggressively toward researching human–ML augmentation, especially the nuances of this construct. The lack of nuanced developments in augmentation may result in ineffective theories that do not do justice to the complexity of human–ML partnerships in practice, resulting in superficial and flawed understandings of issues of critical importance to organizations and society. The IS discipline's unique perspective uniquely suits the ethical problems in new technology adoption for organizations. We call for IS researchers to engage in more interdisciplinary scholarship to help others understand this rapidly evolving landscape. Consistent with its long history, we also believe that the IS discipline is well suited to exploring different variants of augmentation, including their respective limitations, boundary conditions, and unintended consequences. We hope that this paper will motivate other researchers to explore these research questions and uncover more critical nuances of augmented partnerships. A robust research agenda regarding fairness and augmentation can help organizations more effectively leverage ML models' benefits while limiting the potential adverse societal effects.

Acknowledgments

The authors thank USAID Grant AID-OAA-A-12-00095, "Appropriate Use of Machine Learning in Developing Country Contexts," which partially funded this project as part of Massachusetts Institute of Technology D-Lab Comprehensive Initiative for Technology Evaluation (CITE) and the Carroll School of Management at Boston College for research funding. Mike Teodorescu was supported during part of this research via a visiting scholar appointment at MIT D-Lab. All authors are grateful to the editors of the MISQ Special Issue on Managing AI, the two anonymous referees, and the reviewers of the Academy of Management 2020, Strategic Management Society 2020, the MISQ Special Issue Paper Development Workshop 2020, the NYU AI Workshop 2020, and the Society for Business Ethics 2020 Annual Meeting. The authors are especially grateful for the guidance of senior editor Nicholas

Berente, who provided essential guidance through the review process. The authors also thank Daniel Frey, Kendra Leith, Nancy Adams, Amit Gandhi (all MIT D-Lab), Aubra Anthony and Schachee Doshi of USAID), Sam Ransbotham and Robert Fichman of Boston College, and John Deighton of Harvard Business School for valuable feedback.

References

Aghion, P., Jones, B. F., and Jones, C. I. 2017. "Artificial Intelligence and Economic Growth," Working Paper 23928, National Bureau of Economic Research.

Agrawal, A., Gans, J., and Goldfarb, A. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence*, Boston: Harvard Business Press.

Altman, N., and Krzywinski, M. 2018. "The Curse(s) of Dimensionality," *Nature Methods* (15:6), pp. 399-400.

Ambrose, M. L., and Schminke, M. 2009. "The Role of Overall Justice Judgments in Organizational Justice Research: A Test of Mediation," *Journal of Applied Psychology* (94:2), pp. 491-500.

August, T., and Tunca, T. I. 2006. "Network Software Security and User Incentives," *Management Science* (52:11), pp. 1703-1720.

Arora, A., Caulkins, J. P., Telang, R. 2006. "Research Note: Sell First, Fix Later: Impact of Patching on Software Quality," *Management Science* (52:3), pp. 465-471.

Awwad, Y., Fletcher, R., Frey, D., Gandhi, A., Najafian, M., and Teodorescu, M. 2020. "Exploring Fairness in Machine Learning for International Development," MIT D-Lab, Comprehensive Initiative on Technology Evaluation, Massachusetts Institute of Technology (https://dspace.mit.edu/handle/1721.1/126854).

Barnes, P. 2019. "Artificial Intelligence Poses New Threat to Equal Employment Opportunity," *Forbes*, November 10 (https://www.forbes.com/sites/patriciagbarnes/2019/11/10/artificial-intelligence-poses-new-threat-to-equal-employment-opportunity/?sh=6f6b1e6b6488).

Bazerman, M. H., and Tenbrunsel, A. E. 2012. *Blind Spots: Why We Fail to Do What's Right and What to Do about It*, Princeton, NJ: Princeton University Press.

Benner, K., Thrush, G., and Isaac, M. 2019. "Facebook Engages in Housing Discrimination with its Ad Practices, U.S. Says," *The New York Times*, Politics, March 28.

Bostrom, R. P., and Heinen, J. S. 1977. "MIS Problems and Failures: A Sociotechnical Perspective. Part I: The Causes," *MIS Quarterly*, pp. 17-32.

Burton-Jones, A, and Grange, C. 2013. "From Use to Effective Use: A Representation Theory Perspective," *Information Systems Research* (24:3), pp. 632-658.

Cavusoglu, H., Cavusoglu, H., and Zhang, J. 2008. "Security Patch Management: Share the Burden or Share the Damage?," *Management Science* (54:4), pp. 657-670.

Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. 2019. "Fairness under Unawareness: Assessing Disparity When Protected Class Is Unobserved," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, pp. 339-348.

Chouldechova, A. 2017. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big Data* (5:2), pp.153-163.

Cooper, A. F., and Abrams, E. 2021. "Emergent Unfairness: Normative Assumptions and Contradictions in Algorithmic Fairness-Accuracy Trade-Off Research," in *Proceedings of the 4th AAAI/ACM Conference on AI, Ethics, and Society*, Virtual Conference, May 19-21.

Cropanzano, R., Goldman, B., and Folger, R. 2003. "Deontic Justice: The Role of Moral Principles in Workplace Fairness," *Journal of Organizational Behavior* (24), pp. 1019-1024.

D'Arcy, J., Hovav, A., and Galletta, D. 2009. "Decision-Maker Awareness of Security Countermeasures and its Impact on Information Systems Misuse: A Deterrence Approach," *Information Systems Research* (20:1), pp. 79-98.

Dastin, J. 2018. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women," *Reuters Business News*, October 10 (https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G).

Daugherty, P., and Wilson, H. J. 2018. *Human + Machine: Reimagining Work in the Age of AI*, Boston: Harvard Business Review Press.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, E., and Quattrociocchi, W. 2016. "The Spreading of Misinformation Online," *Proceedings of the National Academy of Sciences* (113:3), pp. 554-559.

Dinov, I. D. 2016. "Methodological Challenges and Analytic Opportunities for Modeling and Interpreting Big Healthcare Data," *Gigascience* (5:1), pp. 1-15.

Ferneley, E. H., and Sobreperez, P. 2006. "Resist, Comply or Workaround? An Examination of Different Facets of Decision-Maker Engagement with Information Systems," *European Journal of Information Systems* (15:4), pp. 345-356.

Fong, S., Si, Y. W., and Biuk-Aghai, R. P. 2009. "Applying a Hybrid Model of Neural Network and Decision Tree Classifier for Predicting University Admission," in *Proceedings of the 7th International Conference on Information, Communications and Signal Processing*, Washington, DC: IEEE Computer Society Press, pp. 1-5.

Fletcher, R. R., Olubeko, O., Sonthalia, H., Kateera, F., Nkurunziza, T., Ashby, J. L., Riviello, R., and Hedt-Gauthier, B. 2019. "Application of Machine Learning to Prediction of Surgical Site Infection," in *Proceedings of the 42st Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* Washington, DC: IEEE Computer Society Press, pp. 2234-2237.

Frank, D-A., Chrysochou, P., Mitkidis, P., and Ariely, D. 2019. "Human Decision-Making Biases in the Moral Dilemmas of Autonomous Vehicles," *Scientific Reports* (9), Article 13080 (https://doi.org/10.1038/s41598-019-49411-7).

Frenkel, S., Alba, D., and Zhong, R. 2020. "Surge of Virus Misinformation Stumps Facebook and Twitter," *The New York Times*, Technology, March 8, (https://www.nytimes.com/2020/03/08/technology/coronavirus-misinformation-social-media.html).

Glikson, E., and Woolley, A. W. 2020. "Human Trust in Artificial Intelligence: Review of Empirical Research," *Academy of Management Annals* (14:2), pp. 627-660.

Goodhue, D. L., and Thompson, R. L. 1995. "Task–Technology Fit and Individual Performance," *MIS Quarterly* (19:2), pp. 213-236.

Gundlach, G. T., and Cadotte, E. R. 1994. "Exchange Interdependence and Interfirm Interaction: Research in a Simulated Channel Setting," *Journal of Marketing Research* (31:4), pp. 516-532.

Hao, K. 2019. "AI Is Sending People to Jail—And Getting it Wrong," *MIT Technology Review*, January 21.

Hao, K., and O'Neill, P. H. 2020. "The Hack That Could Make Face Recognition Think Someone Else Is You," *MIT Technology Review*, August 5, 2020.

Hao, K., and Stray, J. 2019. "Can You Make AI Fairer than a Judge? Play Our Courtroom Algorithm Game," *MIT Technology Review*, October 17, 2019.

Hardt, M., Price, E., and Srebro, N. 2016. "Equality of Opportunity in Supervised Learning," *Advances in Neural Information Processing Systems 29,* D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), pp. 3315-3323.

Imana, B., Korolova, A., and Heidemann, J. 2021. "Auditing for Discrimination in Algorithms Delivering Job Ads," in *Proceedings of International World Wide Web Conference*, April 2021, Ljubljana, Slovenia.

Jonas, E., Schulz-Hardt, S., Frey, D., and Thelen, N. 2001. "Confirmation Bias in Sequential Information Search after Preliminary Decisions: An Expansion of Dissonance Theoretical Research on Selective Exposure to Information," *Journal of Personality and Social Psychology* (80:4), pp. 557-571.

Kane, G. C., and Alavi, M. 2008. "Casting the Net: A Multimodal Network Perspective on Decision-Maker–System Interactions," *Information Systems Research* (19:3), pp. 253-272.

Kane, G. C., Alavi, M., Labianca, G., and Borgatti, S. P. 2014. "What's Different about Social Media Networks? A Framework and Research Agenda," *MIS Quarterly* (38:1), pp. 275-304.

Kane, G. C., Palmer, D., Phillips, A. N. 2019. "Accelerating Digital Innovation Inside and Out: Agile Teams, Ecosystems, and Ethics," *MIT Sloan Management Review* (https://sloanreview.mit.edu/projects/accelerating-digital-innovation-inside-and-out).

Kane, G. C., Young, A., Majchrzak, A., and Ransbotham, S. 2021. "Avoiding an Oppressive Future: Designing Emancipatory Autonomous Agents," *MIS Quarterly* (45:1), pp. 371-396.

Kauffman, S. A., and Weinberger, E. D. 1989. "The NK Model of Rugged Fitness Landscapes and its Application to Maturation of the Immune Response," *Journal of Theoretical Biology* (141:2), pp. 211-245.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. 2018. "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause (eds.), Stockholm, pp. 2564-2572.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. 2017. "Avoiding Discrimination Through Causal Reasoning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 656-666.

Kleinberg, J., Mullainathan, S., and Raghavan, M. 2016. "Inherent Tradeoffs in the Fair Determination of Risk Scores," in *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, C. H. Papadimitrou (ed.), Article 43, pp. 43:1-43:23.

Kobayashi, M., Fussell, S. R., Xiao, Y., and Seagull, F. J. 2005. "Work Coordination, Workflow, and Workarounds in a Medical Context," in *Proceedings of CHI 2005 Extended Abstracts on Human Factors in Computing Systems*, Portland, OR, pp. 1561-1564.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. 2017. "Counterfactual Fairness," in *Proceedings of the 31st Conference on Neural Information Processing Systems*, Long Beach, CA, pp. 4066-4076.

Lee, D. J., Ahn, J. H., and Bang, Y. 2011. "Managing Consumer Privacy Concerns in Personalization: A Strategic Analysis of Privacy Protection," *MIS Quarterly* (35:2), 423-444.

Leonardi, P. M. 2011. "When Flexible Routines Meet Flexible Technologies: Affordance, Constraint, and the Imbrication of Human and Material Agencies," *MIS Quarterly* (35:1), pp. 147-167.

Levy, P. E., and Williams, J. R. 2004. "The Social Context of Performance Appraisal: A Review and Framework for the Future," *Journal of Management* (30:6), pp. 881-905.

Lindebaum, D., Vesa, M., and den Hond, F. 2020. "Insights From the Machine Stops to Better Understand Rational Assumptions in Algorithmic Decision Making and its Implications for Organizations," *Academy of Management Review* (45:1), pp. 247-263.

Lu, R., Chatman, J. A., Goldberg, A., and Srivastava, S. B. 2019. "Situated Cultural Fit: Value Congruence, Perceptual Accuracy, and the Interpersonal Transmission of Culture," Working Paper, University of California, Berkeley.

Martin, K. 2019. "Ethical Implications and Accountability of Algorithms," *Journal of Business Ethics* (160:4), pp. 835-850.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. 2019. "A Survey on Bias and Fairness in Machine Learning," arXiv:1908.09635.

Meyer, D. 2018. "Amazon Reportedly Killed an AI Recruitment System Because it Couldn't Stop the Tool from Discriminating Against Women," *Fortune,* October 10 (https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/).

Meyer, M. H., and Curley, K. F. 1991. "An Applied Framework for Classifying the Complexity of Knowledge-Based Systems," *MIS Quarterly* (15:4), pp. 455-472.

Milkman, K. L., Rogers, T., and Bazerman, M. H. 2008. "Harnessing Our Inner Angels and Demons: What We Have Learned about Want/should Conflicts and How That Knowledge Can Help Us Reduce Short-Sighted Decision Making," *Perspectives on Psychological Science* (3:4), pp. 324-338.

Mitchell, T. M. 1997. *Machine Learning,* New York: McGraw-Hill Science.

Murray, A., Rhymer, J., and Sirmon, D. G. 2021. "Humans and Technology: Forms of Conjoined Agency in Organizations," *Academy of Management Review* (forthcoming) (https://doi.org/10.5465/amr.2019.0186).

Nan, N. 2011. "Capturing Bottom-Up Information Technology Use Processes: A Complex Adaptive System Model," *MIS Quarterly* (35:2), pp. 505-532.

O'Donnell, S. W. 2000. "Managing Foreign Subsidiaries: Agents of Headquarters, or an Interdependent Network?," *Strategic Management Journal* (21:5), pp. 525-548.

Orlikowski, W. J. 2000. "Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations," *Organization Science* (11:4), pp. 404-428.

Orlikowski, W. J., and Scott, S. V. 2008. "10 Sociomateriality: Challenging the Separation of Technology, Work and Organization," *Academy of Management Annals* (2:1), pp. 433-474.

Pich, M. T., Loch, C. H., and Meyer, A. D. 2002. "On Uncertainty, Ambiguity, and Complexity in Project Management," *Management Science* (48:8), pp. 1008-1023.

Raisch, S., and Karkowski, S. 2020. "Artificial Intelligence and Management: The Automation–Augmentation Paradox," *Academy of Management Review* (Forthcoming), (doi: 10.5465/2018.0072).

Recker, J. 2022. *Scientific Research in Information Systems: A Beginner's Guide* (2nd ed.), Berlin: Springer.

Rivera, L. A. 2012. "Hiring as Cultural Matching: The Case of Elite Professional Service Firms," *American Sociological Review* (77:6), pp. 999-1022.

Saravanakumar, K. K. 2021. "The Impossibility Theorem of Machine Fairness—A Causal Perspective," arXiv:2007.06024.

Shore, J., Baek, J., and Dellarocas, C. 2018. "Network Structure and Patterns of Information Diversity on Twitter," *MIS Quarterly*. (42:3), pp. 849-872.

Smith-Crowe, K., Tenbrunsel, A. E., Chan-Serafin, S., Brief, A. P., Umphress, E. E., and Joseph, J. 2015. "The Ethics 'Fix"' When Formal Systems Make a Difference," *Journal of Business Ethics* (131:4), pp. 791-801.

Sommer, S. C., and Loch, C. H. 2004. "Selectionism and Learning in Projects with Complexity and Unforeseeable Uncertainty," *Management Science* (50:10), pp. 1334-1347.

Tjosvold, D. 1986. "The Dynamics of Interdependence in Organizations," *Human Relations* (39:6), pp. 517-540.

ur Rehman, M. H., Liew, C. S., Abbas, A. Jayaraman, P. P., Wah, T. Y., and Khan, S. U. 2016. "Big Data Reduction Methods: A Survey," *Data Science and Engineering* (1), pp. 265-284,

Van Huynh, T., Van Nguyen, K., Nguyen, N. L. T., and Nguyen, A. G. T. 2020. "Job Prediction: From Deep Neural Network Models to Applications," in *Proceedings of the 2020 RIVF International Conference on Computing and Communication Technologies*, Washington, DC: IEEE Computer Society Press, pp. 1-6.

van den Bos, K., and Lind, E. A. 2002. "Uncertainty Management by Means of Fairness Judgments," in *Advances in Experimental Social Psychology* (Vol. 34), M. P. Zanna (ed.), New York: Academic Press, pp. 1-60.

Wang, W. Q., and Benbasat, I. 2007. "Recommendation Agents for Electronic Commerce: Effects of Explanation Facilities on Trusting Beliefs," *Journal of Management Information Systems* (23:4), pp. 217-246.

Watson, R. T., DeSanctis, G., and Poole, M. S. 1988. "Using a GDSS to Facilitate Group Consensus: Some Intended and Unintended Consequences," *MIS Quarterly* (12:3), pp. 463-478.

Wick, M., Panda, S., Tristan, J. P. 2019. "Unlocking Fairness: A Trade-Off Revisited,"in *Proceedings of the 33rd Conference on Neural Information Processing Systems,* Vancouver, Canada.

Winkler, J. K., Sies, K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., Abassi, M., Fuchs, T., Rosenberger, A., and Haenssle, H. A. 2020. "Melanoma Recognition by a Deep Learning Convolutional Neural Network—Performance in Different Melanoma Subtypes and Localisations," *European Journal of Cancer* (127), pp. 21-29.

Wolf, M. J., Miller, K. W., and Grodzinsky, F. S. 2017. "Why We Should Have Seen That Coming: Comments on Microsoft's Tay 'Experiment,' and Wider Implications," *The ORBIT Journal* (1:2), pp. 1-12.

Xiao, B., and Benbasat, I. 2007. "E-Commerce Product Recommendation Agents: Use, Characteristics, and Impact," *MIS Quarterly* (31:1), pp. 137-209.

Ye, L. R., and Johnson, P. E. 1995. "The Impact of Explanation Facilities on Decision-Maker Acceptance of Expert Systems Advice," *MIS Quarterly* (19:2), pp. 157-172.

Yukl, G., and Michel, J. W. 2006. "Proactive Influence Tactics and Leader Member Exchange," in *Power and Influence in Organizations: New Empirical and Theoretical Perspectives*, C. A. Schriesheim and L. L. Neider (eds), Greenwich, CT: Information Age Publishing, pp. 87-103.

About the Authors

Mike H. M. Teodorescu is an assistant professor of Information Systems at Boston College's Carroll School of Management. He received his doctorate from Harvard Business School in Strategy and his bachelor's in computer science from Harvard College. Part of the research in this paper was funded by a Visiting Scholar position at Massachusetts Institute of Technology D-Lab via USAID. His research focuses on machine learning and AI fairness, innovation strategy in firms, and patent policy, integrating analytical rigor and pragmatic insights gained from his wide-ranging experiences in the tech industry and public sector. His research has been published in journals such as *Strategic Entrepreneurship Journal*, *Military Medicine*, *SAE Transactions*, as well as proceedings such as the Academy of Management Best Paper Proceedings, International Conference on Information Systems, and various IEEE conferences. His work has been featured in HBS Working Knowledge, *National Defense* (NDIA's business and technology magazine), TEDx London xLab, the London Design Museum, NECN, and other media.

Lily Morse is an assistant professor of Management at the John Chambers College of Business & Economics, West Virginia University. Her research seeks to understand how organizations can reduce unethical behavior in the workplace, which she has investigated in the context of negotiation, public auditing, and counter-

productive workplace behaviors. Her latest research examines how AI systems can be made more capable of making moral decisions that improve people's experiences of fairness at work. She has published in journals such as *Organizational Behavior and Human Decision Processes, Academy of Management Perspectives*, *Journal of Personality and Social Psychology, Journal of Research in Organizational Behavior*, and *Journal of Research in Personality*. Her research has been covered by media outlets such as Business Insider, Scientific American, Huffington Post, and New Scientist. She received her Ph.D. in Organizational Behavior and Theory from Carnegie Mellon University.

Yazeed Awwad is a research associate at the Center for Complex Systems at KACST and MIT, with a focus on complex networks of interconnected information. He received his M.S. from the Massachusetts Institute of Technology. While an affiliate at MIT's D-Lab, Yazeed developed a framework for identifying unfair application of ML algorithms in a general sense without restricting fairness to specific protected attributes. While mainly from a quantitative and technical background, Yazeed presented his MIT thesis work, a study of the emergence and evolution of concepts represented via Wikipedia article links, at the 31st Annual Convention of the Association for Psychological Sciences in 2019 as part of a symposium with Professor David Dunning.

Gerald C. (Jerry) Kane is a professor of Information Systems and Faculty Director of the Edmund H. Shea, Jr. Center for Entrepreneurship at Boston College's Carroll School of Management. He researches and teaches about how companies can understand and respond to digital disruption to undergraduate, graduate, and executive education students worldwide. He has published over 100 papers, articles, and reports on these topics in journals such as *MIS Quarterly, Information Systems Research, Organization Science, Management Science, Journal of Management Information Systems, Harvard Business Review,* and *MIT-Sloan Management Review*, among others. He is lead author of *The Technology Fallacy: How People are the Real Key to Digital Transformation* (MIT Press). He is publishing a follow-up book in September 2021, entitled *The Transformation Myth: Leading Your Organization Through Uncertain Times* (MIT Press), that focuses on how companies innovate in response to disruptions like COVID-19.