The Best of Both Worlds: Machine Learning and Behavioral Science in Operations Management

Andrew M. Davis[a] • Shawn Mankad[b] • Charles J. Corbett[c] • Elena Katok[d]

adavis@cornell.edu • smankad@ncsu.edu • charles.corbett@anderson.ucla.edu • ekatok@utdallas.edu

[a] *Samuel Curtis Johnson Graduate School of Management, SC Johnson College of Business, Cornell University, Ithaca, NY, 14853*
[b] *Poole College of Management, NC State University, Raleigh, NC 27695*
[c] *Anderson School of Management, University of California, Los Angeles, CA, 90095*
[d] *Naveen Jindal School of Management, University of Texas at Dallas, Richardson, TX 75080*

Abstract. ***Problem definition:*** Two disciplines increasingly applied in operations management (OM) are machine learning (ML) and behavioral science (BSci). Rather than treating these as mutually exclusive fields, we discuss how they can work as complements to solve important OM problems. ***Methodology/results****:* We illustrate how ML and BSci enhance one another in non-OM domains before detailing how each step of their respective processes can benefit the other in an OM setting. We then conclude by proposing a framework to help identify how ML and BSci can jointly contribute to OM problems. ***Managerial implications****:* Overall, we aim to explore how the integration of ML and BSci can enable researchers to solve a wide range of problems within OM, allowing future research to generate valuable insights for managers, companies, and society.

Keywords: Operations management, machine learning, behavioral science.

1. Introduction

Machine learning (ML), which often relies on big data and deep learning, and behavioral science (BSci), which relies on human-subject experiments and behavioral modeling, offer great potential for operations management (OM) research. Typically, ML utilizes extensive data to improve decision-making and prediction, but it is not without its drawbacks, requiring vast amounts of data and raising concerns about opacity and accountability. These can be partially remedied by drawing on BSci techniques to understand how human decision-makers respond to various factors, often through controlled experiments that generate their own data. ML techniques can be used to work with large datasets capturing aspects of human behavior, such as text or facial cues, that would have previously been inaccessible to BSci, expanding the available research toolkit. We describe how ML and BSci can work as complements to solve important OM problems. The disciplinary separation of ML and BSci scholars means that the benefits of this complementarity are not automatic; we offer this article to highlight the value of forging closer connections for both communities.

One simplistic description of ML is that it is a form of engineering that aims to solve problems, whereas BSci is a science that aims to understand how the world works. Engineering- and science-oriented approaches naturally drift apart over time (Corbett and Van Wassenhove 1993); even within OM, ML and BSci are often treated as mutually exclusive in their methodologies and results. However, there is

significant overlap. Standard ML techniques can be used to analyze data and develop algorithms to better predict human decisions (Mišić and Perakis 2022), sometimes allowing causal inference. BSci's behavioral analytical modeling and controlled human-subject experiments can generate causal inferences for predicting decisions. In this sense, in the OM field, the objectives of ML and BSci intersect.

Given this overlap, there is value in leveraging both ML and BSci when examining OM problems, which often involve humans as managers, employees, and consumers (Buell 2018, Roels and Staats 2021). Despite companies' increased use of recommender systems, these are often overridden by the human managers entrusted with final decisions (Siemsen and Aloysius 2020). ML can be problematic when unattended, but managers making complex decisions without automated support can also have adverse outcomes.

We adopt broad definitions of ML and BSci and employ methods and results from the respective literatures. We use ML in its most modern incarnation, which includes tools for big data, deep learning, artificial intelligence, and other techniques for inductive prediction. Also included are unsupervised and supervised learning algorithms; the latter allows access to annotated or labeled data for learning and prediction. Relevant results in the field of ML include findings around how model performance and complexity and the volume and dimension of data relate (e.g., overfitting, the bias-variance tradeoff, and the curse of dimensionality). Similarly, we draw on BSci's human-subject experiments and behavioral analytical models, along with established results, including those concerning the impact of behavioral bias on managerial decisions and interventions to improve them.

Challenges faced in ML include data availability and incorporating behavioral factors into algorithms; as Luca et al. (2016) note in "Algorithms Need Managers, Too," it is important to understand what algorithms do well and to recognize their weaknesses. The methods of BSci are useful here. Human-subject experiments can generate clean data to compare the performance of ML algorithms in a relatively risk-free environment (Bastani et al. 2022), and behavioral modeling can prescribe how the input of a human manager can best be balanced with an algorithm (de Véricourt and Gurkan 2023). BSci results can inform initial algorithm design, for example, concerning the behavioral parameters or features to include in the objective function while also ensuring equity, privacy, and transparency (Hunt 2021). In turn, BSci in OM faces problems in identifying hypotheses in unstructured settings and conducting "exploratory" studies. ML can ease this concern, and researchers can, for example, employ text analysis to inform experimental hypotheses for unstructured problems or those lacking clear predictions. Causal forests allow the analysis of large behavioral data sets to identify relevant behavioral hypotheses.

Despite the potential benefits of ML and BSci, few OM studies leverage both successfully (we detail exceptions to this statement below), although this is changing. We can only speculate why this is the case, but some possibilities include the pressure on researchers to have a narrow disciplinary focus. ML and BSci researchers have a different focus in graduate school and, over time, drift further apart. Corrective action is necessary to support ML–BSci collaborations in the future.

The remainder of this paper is organized as follows. In Section 2, we detail how ML and BSci have served as complements, primarily focusing on other fields. We then take a detailed bottom-up approach and consider ML and BSci in the OM context in Section 3. In Section 4, we offer examples of OM studies using ML and BSci together to yield interesting insights. In Section 5, we conclude with a framework for combining ML and BSci based on the objective of the researcher and the availability of data; we consider future research opportunities, the role of graduate student training, and resources for researchers.

2. Illustrating How ML and BSci Can Enhance Each Other in Other Fields

Here we outline how ML and BSci are applied to fields outside of OM. We first detail key strengths of ML, followed by weaknesses. We then do the same for BSci, discussing notable strengths and then weaknesses. We subsequently use this to motivate how the two fields can serve as complements.

2.1 ML in Other Fields

Researchers across fields have begun to investigate how human decision-makers can leverage ML. In healthcare, Rajkomar et al. (2019) examine how ML assists clinicians by identifying symptom patterns, increasing the accuracy and efficiency of diagnosis and prognosis. In finance, ML support systems are so ubiquitous that there are survey papers summarizing the literature; Goodell et al. (2021) note the use of ML for portfolio construction and valuation, forecasting, and addressing fraud and distress. Horton (2017) analyzes hiring recommendation systems in organizational behavior, and Kleinberg et al. (2018) explore how judges can use ML to make better bail decisions through more accurate behavioral forecasting. Other applications include university admissions, eligibility for government subsidies, insurance rates, and fraud detection (Kelliher 2021, West 2021, Uzialko 2022).

Examples abound of how ML can improve decision-making, but it has its challenges and limitations. These generally relate to the availability of high-quality and unbiased data – without which algorithms will make inaccurate predictions – and proper human oversight and training of ML systems – which can become destructive when unattended or when users are unfamiliar with their assumptions (O'Neil 2016). Blythe (2018) notes parallels to financial engineering before the 2008 financial crisis; bigger data sets and greater computing power are now available, but the increasing algorithmic complexity makes it more

challenging to identify factors implicated in previous crises, such as flawed assumptions, gaps in logic, and differences between conditional and actual probabilities.

In "When Machine Learning Goes Off the Rails," Babic et al. (2021) highlight the risks of concept drift (when inputs are not stable over time or are misspecified) and covariate shift (when the training data differ from the application data). Larson (2021) describes the challenges faced by IBM's Watson, a system famous for defeating human champions in the game show *Jeopardy!*; the system exploited a key structural feature specific to the show, namely, that 95% of all answers are Wikipedia titles, allowing relatively simple matching algorithms to have extraordinary gameplay. Watson remains impressive, but the IBM subdivision recently exited the healthcare domain because the structures of real-world problems are too complex to fully automate without significant risk.

Many articles warning of the risks of ML offer ways to mitigate these. Blythe (2018) advocates for the integration of human judgment: "More than ever, judgment – necessarily subjective and based on experience – will play a significant role in moderating over-reliance on and misuse of quantitative models" (para. 13). Babic et al. (2021) recommend executives "treat machine learning as if it's human" (para. 26). Of course, blindly applying human judgment may simply perpetuate bias (De-Arteaga et al. 2022), and BSci offers ways to ensure that human judgment enhances ML.

Kleinberg et al. (2023a) note ML's "inversion problem;" that is, ML infers preferences from data on choices. If an AI trained on a radiologist's readings of images does not know if those readings were impacted by, for example, the radiologist's tiredness or other behavioral factors (such as the sequence of preceding readings), the AI will encode the errors made by the radiologist rather than improving on or avoiding these. The authors call for the development of a "science of inversion," including multiple measures being used to triangulate behavior and performing an inversion on small datasets to help build algorithms for larger data sets. Agan et al. (2023) demonstrate that the more automated the choice behavior, the greater the divergence between choices and preferences and the risk of in-group bias; the authors compare user responses to Facebook's suggestions in its Newsfeed (more automatic) and People You May Know (more deliberate) features. Kleinberg et al. (2023b) develop an analytical model of users' engagement with a platform as driven by a mix of their system 1 (automatic) and system 2 (deliberate) thinking to demonstrate that a platform aiming to maximize engagement cannot use observed behavior to explore user utility. More generally, the context of human choices is critical when using those choices to make predictions. Without behavioral modeling, the algorithm is lost.

2.2 BSci in Other Fields

BSci covers various domains, including psychology (e.g., Kahneman 2011), economics (e.g., Kagel and Roth 2016), marketing (e.g., Bearden and Netemeyer 1999), and accounting (e.g., Libby et al. 2002). BSci often involves prescribing how decisions should be made in complex environments and taking account of potential behavioral bias, usually through analytical modeling (e.g., Boyaci et al. 2023), and its techniques are used to make causal inferences through controlled human-subject experiments. ML methods can also be used for causal inference, but BSci differs in its focus on the behavioral mechanisms underlying human decisions. For example, acting according to ideas of fairness, a decision-maker may deviate from the normative profit-maximizing decision if they are concerned about inequitable outcomes. Fairness may come into play during negotiations over financial contracts all the way to consumers comparing prices they paid for a product, before and after a price promotion is applied. The degree to which a party weighs fairness can even differ depending on whether one is matched with a human or an automated agent, which has implications for algorithm design (Johnsen et al. 2019).

Just as unmonitored algorithms can be dangerous, relying on human judgment can also be risky. Many decisions involve myriad factors that an individual may struggle to incorporate and optimize. For example, without access to dynamic models, which can recommend (or automate) prices for airline tickets and hotel rooms in real-time, a manager would struggle to maximize profits. BSci methods are not without drawbacks. There are often questions about whether the results observed in the controlled environment of a laboratory experiment are generalizable (external validity);[1] laboratory experiments often rely on students or other common-population participants to play the role of managers, and they may not accurately reflect managerial behavior, and the lab setting may not accurately capture managers' work environment. Another weakness is that BSci methods may not produce precise predictions.

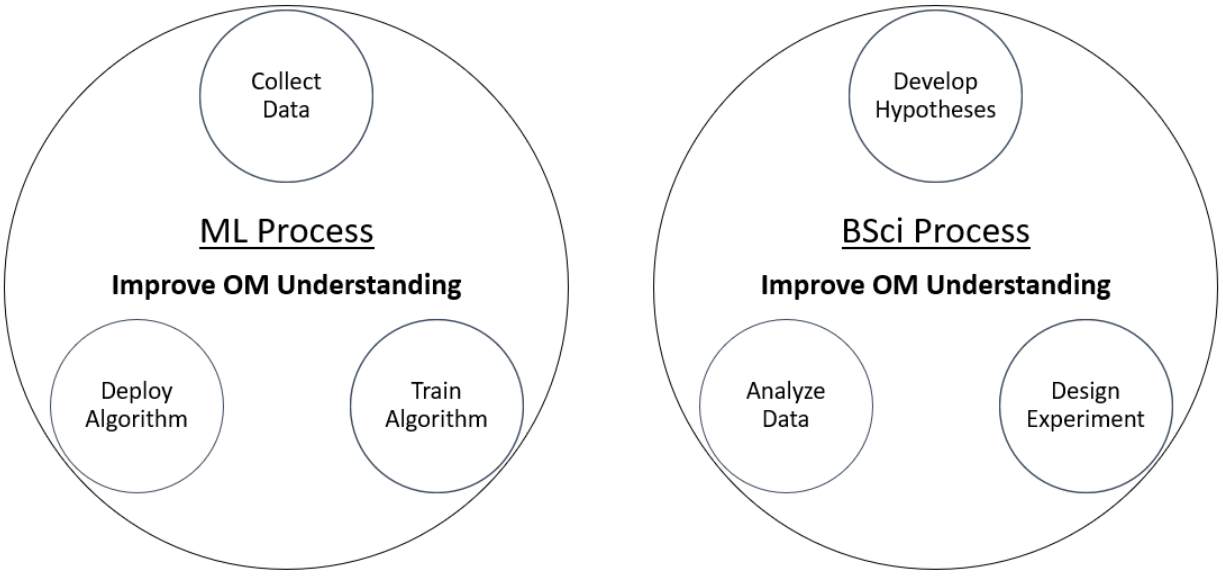
2.3 The Case for ML and BSci Together

The respective weaknesses of ML and BSci offer opportunities for these to be used as complements. ML struggles when high-quality data are lacking or when the ML system may not accurately reflect the preferences of human decision-makers. This is precisely where BSci is most useful: experiments can generate data and help identify the underpinnings of human decisions. Further, BSci analytical modeling can prescribe how human managers, who may be susceptible to a particular behavioral bias, should make decisions employing ML products and systems without requiring data whatsoever (Gurkan and de Véricourt 2022, Chen et al. 2022, please see Cukier et al. 2021 for more general information around modeling in the time of ML). Conversely, external validity, prediction, and high resource demands present challenges to BSci, opening space for ML methods to flourish; these do not require significant lab resources and can be applied to a variety of complex data sets to improve predictive performance and

can help researchers analyze data to identify potential "natural" experiments in the field (e.g., a significant shock).

3. Illustrating How ML and BSci Can Enhance Each Other in OM: Process

In this section, we explore the synergies of ML and BSci in more detail in the context of OM by first considering the contribution of each. We then describe the process by which the ML/BSci synergies can be accessed, providing examples of OM studies that have done so successfully. Figures 1(a) and 1(b) are high-level representations of key steps in ML and BSci research, with connecting arrows deliberately omitted, as the sequence can vary and can be iterative. The circles represent steps for which the alternative field can assist; not every project will include all steps.



Panel (a) ML Process **Panel (b)** BSci Process

Figure 1: Sample ML and BSci Research Processes in OM

3.1 How the ML Research Process Can Benefit From BSci

We first examine how BSci can add value to ML research, following the stages in Figure 1 (a); we list BSci resources for interested ML researchers in Appendix A.

3.1.1 Collect Data. Most ML algorithms rely on large data sets, but such data may not be available or reliable. Internal firm data may not be publicly disclosed or may have censored or missing values. The human-subject experiments and independent data collection of BSci can be used in training and refining

an ML system. This is a natural evolution of the use of Amazon Mechanical Turk (MTurk) and other services in ML for labeling images or text.

While human-subject experiments yield fewer observations than ML methods typically utilize (e.g., hundreds of independent participants in an experimental study), the value of an initial dataset can be significant when prototyping an algorithm. Consider the case of Blue River Technology, which developed an image-recognition algorithm to distinguish between weeds and crops, increasing farmers' efficiency and reducing costs (Ng 2018, Trautman 2018). The initial algorithm performed poorly but was improved by early user feedback, which is consistent with Agrawal et al. (2018), who stress the importance of capturing the accuracy of predictions to continuously improve an algorithm's performance. In 2017, Blue River was sold to John Deere for $305 million (Tobe 2017). In another example, at its launch, Google's voice assistant was initiated by saying, "Hey Google." Because this phrase was contrived, Google created the initial training data by recording individuals saying this and other phrases in different contexts (i.e., a controlled experiment was used to generate the initial training data). Once deployed, the underlying algorithms could be refined over time with real user experience data.

In settings where unlabeled data are abundant, but labels are difficult, time-consuming, or expensive, active learning (Settles 2009) can be employed to allow the ML algorithm to choose the data from which it learns. Specifically, in active learning, the algorithm requests particular unlabeled instances be labeled for the model to be re-estimated. The labeling may be done by human annotators or generated through experiments. For example, suppose a call center aims to improve workforce scheduling. A lab experiment could reveal the effectiveness of different scheduling options, which can be labeled and used to update the ML model parameters.

Tools built on deep learning (LeCun et al. 2015) and large language models (LLMs), such as BERT (Devlin et al. 2019) or Chat-GPT (Brown et al. 2020), have proved popular, and transfer learning has emerged as an important process in the use of ML and AI for specific applications. Transfer learning involves fine-tuning a pre-trained model on a smaller dataset specific to the context or prediction task of interest. For example, an e-commerce company may wish to enhance its customer service using a chatbot. The goal of transfer learning would be to fine-tune the chatbot's responses based on specific behavioral insights, for example, from lab experiments related to customer interactions. The experiment could simulate customer service scenarios with participants playing customers with specific issues or feedback. Data from the lab experiment can reveal how customers respond to different tones, levels of formality, and specific language nuances in the chatbot's responses. The model's parameters can then be adjusted to align with preferred customer interactions, incorporating specific phrases, tone adjustments, and

presentation styles identified in the behavioral analysis; for a technical discussion of how fine-tuning can be implemented, see Niu et al. (2020), Weiss et al. (2016), and Pan and Yang (2010), and references therein.

Table 1. Alternative Types of Human-Subject Experiments and Their Common Characteristics

	Laboratory	**Online/Remote**		**Field**
Mode	In-Person	Synchronous	Asynchronous	Field
Participants	University Students	Students, General Population, Managers		General Population, Managers
Internal Control	High	High-to-Medium	Medium-to-Low	Low
External Validity	Low	Moderate		High

How does one go about administering experiments to generate data? Fortunately, there are excellent resources in this area, and we offer only a brief overview; see Appendix A. Table 1 illustrates the modes and participant types for various experiment types, along with their internal control and external validity tradeoffs. Laboratory experiments typically involve university students and are administered in person, offering high internal control. The data-generation process is transparent in that researchers control the environment and observe decision-making. This results in a high signal-to-noise ratio for data, often with no missing values and balanced panels. Researchers can also design the experiment to observe "tail" events (low-probability events that are difficult to identify in practice, such as extreme shocks to a supply chain).

External validity – whether the observed results are generalizable to practice – may be challenging in such experiments. While students may make decisions like general consumers and managers (Bolton et al. 2012) in certain laboratory experiments, there is always the possibility that the standard laboratory implementation may not capture other characteristics that exist in practice. As Table 1 shows, there has been a shift toward running experiments online, often with a university's participant pool or platform like MTurk or Prolific (online studies can also be conducted with managers). Online experiments can be synchronous or asynchronous. Synchronous experiments can be administered like an in-person laboratory experiment using a web conferencing application; participants are able to read instructions and answer questions in private, maintaining medium-to-high internal control. In asynchronous online experiments, a researcher emails a link to each participant who completes the task on their own. Internal control is

weaker because participants may not be focused on the task, or a bot may be making the decisions; periodic attention checks, removing duplicate IP addresses, or requiring a minimum performance rating of participants can be used to avoid such challenges. As set out in Table 1, field experiments are usually conducted by an organization, allowing the researcher to implement the necessary treatments with employees or consumers. Challenging to implement and allowing less control since they are essentially conducted "in the wild," field experiments do have the benefit of high external validity; for more on best practices in OM field experiments, see Ibanez and Staats (2019).

For many applications, ML algorithms must account for social preferences and societal norms. To the extent that data come from self-interested individuals in the course of normal transactions, they may reflect unethical behavior (such as discrimination, selfishness, or bullying). Algorithms trained using these data will exhibit the same undesirable features, and it is difficult to attribute responsibility for the actions of algorithms (Pasquale 2015). Increasing algorithmic transparency is often cited as a potential solution. However, it is not well understood when or how people will incorporate the fact that their behavior, by inadvertently training algorithms, may affect others; data collected through BSci experiments can alleviate this problem. For example, Klockmann et al. (2022) show that merely informing lab participants that their actions are used to train an algorithm that will make decisions with monetary consequences does not change behavior. However, an explicit understanding of the risk of being harmed by the decisions of unethical algorithms reduces selfishness, an insight made possible by collecting data using BSci methods.

In all cases, being mindful of the distinction between choices and preferences is essential. As Kleinberg et al. (2023a) point out, individuals often make choices inconsistent with their preferences. Applying ML to observed choices will often increase engagement but decrease user utility. The same risk applies to OM models applying ML to optimize product lines by observing purchase behavior, especially for items more likely to be impulse buys. For instance, it is conceivable that some customers of fast fashion brand Shein are seduced into buying items that, if they took the time to consider the unsustainable nature of their purchases, they would realize they do not really want. Using BSci to infer preferences from observed choices, as Kleinberg et al. (2023a) advocate, reduces that risk.

3.1.2 Train Algorithm. It is increasingly possible for researchers to consider behavioral factors when training ML algorithms (by "training," we refer to the initial design and early refinements of an algorithm in addition to estimation). Results from BSci can be included as features of the ML model, in refining model parameters, or even as a component of the objective function (i.e., likelihood function) defining the ML model to "nudge" the algorithm and ultimately end-users toward preferable outcomes. Many of the OM-relevant results in the BSci literature pertain to behavioral bias (a non-monetary preference), which can

be observed in individual decisions and strategic interactions; see Davis (2019) and Bolton and Chen (2019) for details. Biases in individual decisions are often separated into judgments regarding risk and those involving outcomes. The biases of the former type include representativeness, base-rate neglect, anchoring, availability, probability weighting, overconfidence, and ambiguity aversion. Many of these can be corrected through ML systems. For instance, the human tendency to overlook base-rate probabilities when calculating a posterior probability can be avoided with automated systems. BSci can help identify the potential gains of an ML system's quantification of errors.

Outcome-related bias (e.g., risk aversion, prospect theory, anticipated regret, reference dependence, mental accounting, intertemporal choice, and the endowment effect) is especially relevant to ML researchers. Consider prospect theory, a behavioral model that incorporates the idea that losses feel worse than gains feel good (Kahneman and Tversky 1979). If a company uses an algorithm to start charging for a service that was previously free, disgruntled consumers may turn to their competitors. Incorporating such behavioral tendencies can help a company financially and reputationally. Turning to strategic interactions, noteworthy biases established by BSci include fairness and trust. For example, Özer et al. (2011) study a game in which a retailer has private information about demand and can send a signal to the supplier, who then decides on capacity. If the parties only care about profits, the retailer should inflate its forecast, and the supplier should ignore it. However, in experiments, when their private information suggests higher demand, retailers signal this, and suppliers respond by expanding capacity. Last, although not a bias per se, random errors by human decision-makers can also be incorporated to benefit ML algorithms; see Taylor (2016) for a specific example of expected utility "quantilization," where the algorithm randomly selects from a top proportion of actions.

A comparison of game theory and behavioral game theory offers insight into how ML can benefit from BSci in the training of algorithms: "Game theory, the formalized study of strategy, [asks] how emotionless geniuses should play games, but ignored until recently how average people with emotions and limited foresight actually play games" (Camerer 2003). Behavioral game theory incorporates observed biases, for example, those documented in BSci, into the normative game theoretic model (see Cui and Wu 2019 for details). ML researchers can follow a similar approach by tuning a new ML algorithm through different parameters or adding features driven by known biases to ensure an algorithm's upfront accuracy.

Ethics are paramount in the design and training of algorithms, a topic too extensive and important to do justice here. However, we would be remiss if we did not mention that BSci can assist ML with respect to transparency and equity (Hunt 2021), specifically by mitigating the risk of discriminatory bias (e.g., racial). Issues of transparency include algorithmic black boxes and opaqueness and the associated

challenge of attributing responsibility for algorithmic (mis)behavior. Importantly, controlled BSci laboratory experiments can be used to consider the potential impacts of algorithms on human behavior, to identify when an algorithm may be generating discriminatory recommendations and outcomes, and to test potential interventions *before* implementation, thereby avoiding adverse consequences.

3.1.3 Deploy Algorithm. BSci, in particular behavioral theoretical modeling, can shed light on how humans may interact with an algorithm once it is deployed. Recent work in OM employs this methodology to characterize how a human manager, who may have cognitive limitations, makes decisions when presented with an algorithmic recommendation. For instance, Boyaci et al. (2023) develop a "rational inattention framework" in a setting where an algorithm provides a human manager with accurate but incomplete information. They demonstrate that such algorithmic input always improves human decision accuracy but can *increase* false positives, a unique insight that may not have been recognized without the behavioral model. In a related vein, Chen et al. (2022) develop a theoretical model to study how algorithms can be safeguarded by human knowledge (an algorithmic recommendation is altered if it appears unreasonable to a human decision-maker) and the extent to which this guardrail is beneficial.

Algorithms deployed in practice may encounter algorithm aversion (de Véricourt and Perakis 2020), which is the reluctance of human decision-makers to accept recommendations from an automated system, especially when it is a black box. This phenomenon is observed in various OM settings, including retail operations (Caro and de Tejada Cuenca 2018, Kesavan and Kushwaha 2020) and warehousing (Bai et al. 2022, Sun et al. 2022). The formative experiments on algorithm aversion by Dietvorst et al. (2015) show that forecasters lose confidence in an algorithm, particularly after seeing it make the same mistake repeatedly. In fact, in environments characterized by uncertainty, even optimal recommendations are "wrong" ex-post, decreasing the confidence of people who mistakenly assess the quality of a decision by its outcome. BSci methods can yield insights on addressing this risk. For example, using experiments, Bolton and Katok (2018) show that compliance is improved by providing decision-makers with information about the uncertainty of an event, in addition to a recommendation. Other examples include Zhang and Cao (2022), who find that a "forced intervention" leads to greater compliance in the long run.

Burton et al. (2020) review 61 studies on algorithm aversion, finding that most rely on BSci experiments (26 experimental studies, 32 conceptual/review articles, and three ethnographies/field studies). However, BSci theoretical modeling can also be beneficial in this area. For instance, de Véricourt and Gurkan (2022) examine a setting in which a human supervises an algorithm in a high-stakes setting (whether a biopsy should be performed). The human participant observes the performance of the algorithm over time and updates their beliefs about its recommendations but suffers from a "verification

bias" in that they can only update such beliefs when a biopsy is actually performed. De Véricourt and Gurkan (2022) identify conditions under which such learning failures occur and outline suggestions for when to accept or reject automated recommendations. Dai and Singh (2023) consider this problem in a theoretical way in a healthcare setting, exploring two patient-protection and insurance schemes that affect a physician's liability when they have access to an algorithm. The authors show, among other things, that physicians overuse the algorithm for low-uncertainty scenarios when it provides little value and may avoid its use in high-uncertainty scenarios, even though it could have led to better decisions. Note these studies can help show when algorithm aversion may be likely without any data at all.

Turning to the refinement phase of deploying ML algorithms, active and transfer learning from BSci (discussed in Section 3.1.1) can play an important role via controlled lab experiments. However, "debiasing" offers another way to refine algorithms and is frequently used to correct decision-makers who exhibit a bias that is not in their best interest. Ren and Croson (2013), for example, posit that individuals making newsvendor decisions underestimate the variance of the demand distribution, leading to orders between the normative quantity and mean demand (the "pull-to-center" effect). They demonstrate ways to reduce this overconfidence and improve ordering decisions, for example, by having the decision-maker first consider the likelihood of events occurring in the tails of the demand distribution.

How might this work in practice when the order-quantity recommendation might be automated? As a start, the task could be broken into a sequence of smaller steps, and the ML system could present a recommended service level, point forecast, and standard deviation, which may be more acceptable and transparent to a manager than a final recommendation. Such "task decomposition" has been shown to improve decisions in OM contexts (Lee and Siemsen 2017). In a similar vein, research on simple interventions and nudges could be extended to consider how individuals follow ML recommendations; Soll et al. (2015) and Thaler and Sunstein (2009) offer additional details on nudges.

The deployment of ML can be enhanced through BSci, for example, by ensuring that there is proper periodic human oversight of an algorithm's performance to avoid the dangers outlined in Section 2. In many ML deployments, there is no way to turn actual outcomes into new training data, and, as a result, the algorithm never moves from its suboptimal parameter estimates. BSci experiments can generate feedback data when none is otherwise available, and behavioral modeling can be used to prescribe how best to incorporate early usage data (e.g., Gurkan and de Véricourt (2022); details available in Section 4). Finally, even if a researcher incorporates certain behavioral preferences into the algorithm, there may be others that are difficult to quantify, such as happiness or mental health. At a minimum, by monitoring the adverse side effects of the ML system, a researcher can directionally impact some of the algorithm inputs.

3.2 How the BSci Research Process Can Benefit from ML

We next explore how ML can add value to BSci research, following the stages in Figure 1 (b). As before, the sequence can vary, and not every project includes all steps; we provide a list of ML resources for interested BSci researchers in Appendix B.

3.2.1 Develop Hypotheses. When applied to OM, BSci tends to begin with a particular theoretical model. However, for some problems of interest, there is little theory, and behavioral insights could be valuable. For example, infinitely repeated games are useful representations of long-term relationships but often have multiple (or infinite) equilibria. ML techniques can facilitate a more efficient interplay between exploratory and confirmatory research by, identifying potential hypotheses for testing through analysis of supplementary observational unstructured data, such as text and video. For example, participants in one treatment might exhibit a particular behavioral bias detectable through a writing sample solicited as part of the experiment (or a verbal sample, such as the "think aloud" experiments of Gavirneni and Isen (2010)). ML techniques can be used at scale while maintaining rigor and avoiding the time and resources demanded by manual analysis. For instance, the sentiment around dimensions of service quality in customer feedback or online reviews (Mejia et al. 2021), the litigiousness of corporate disclosures (Loughran and McDonald 2011), and the emotion embedded within an exchange (Picard 2000) can be evaluated using standard ML tools, such as topic modeling (Blei et al. 2003, Roberts et al. 2016) or sentiment and emotion analysis (Zhang et al. 2018, Liu 2020). Mullainathan and Rambachan (2023) show how algorithms can generate novel anomalies; their econometric approach could be applied to large datasets of inventory decisions to identify behavioral anomalies, enriching our understanding of how individuals make such decisions and leading to new hypotheses for testing using conventional BSci methods.

Tools built on LLMs, such as BERT (Devlin et al. 2019) or Chat-GPT (Brown et al. 2020), can also be leveraged to frame hypotheses. These tools can perform literature reviews and quickly synthesize large amounts of text, helping researchers identify relevant theories and gaps in the literature. When applied to customer reviews, employee feedback, and other text datasets, researchers and companies can use these tools to identify patterns correlated with behaviors. One could, for instance, formulate hypotheses to understand how behaviors change with task speed or capacity by analyzing employee communication. For image and video analysis, methods from affective computing (Picard 2000) can be useful to BSci research. This growing branch of ML characterizes nonverbal human signs, such as facial expressions, body language, gestures, and tone of voice, to assess emotional states. Modern approaches typically cast the problem as one to be addressed by supervised learning, leveraging advances in deep learning, particularly

deep convolutional neural networks. From an image of a person's face, an algorithm can predict whether they are experiencing anger, disgust, fear, happiness, sadness, surprise, or are feeling neutral. Most emotional picture databases are built on static images or video sequences containing only frontal passport-style faces. Algorithms only trained on these can perform poorly with different head poses, angles, and so on.

Other important developments in ML are open-source software and cloud computing. There are high-quality libraries in R and Python for text and video analysis, though researchers often implement models from scratch. Major cloud service providers (e.g., Amazon Web Services and Microsoft Azure) offer pre-trained models, including sentiment analysis and topic modeling of text and affective computing from image, video, and voice recordings.

Experiments are constrained by the participant pool, budget, time, and other resources. ML techniques can facilitate the efficient use of resources by eliminating potential causal relationships not supported in supplementary observational data. Probabilistic graphical models, notably Bayesian networks or Markov equivalence classes, are techniques by which a set of variables and their conditional dependencies can be represented via a directed graph. Such networks are ideal for taking a past event and predicting the likelihood that any one of several possible causes was a contributing factor. For instance, the relationships between the myriad factors potentially influencing a patron's satisfaction in a service context can be represented using a probabilistic graphical model, which can then be analyzed to identify a small set of potential causal relations for further exploration in an experiment (see Buell et al. 2017). Eberhardt et al. (2023) develop and use such a procedure to reveal the precise causal pathways that undermine instrument validity when measuring the effect of education on income.

ML techniques can also be used to analyze archival or field data to identify unexpected behavior, such as tax aversion in the sharing economy (Cui and Davis 2022). In addition, where a BSci researcher feels that deception (misleading participants) would be necessary in the course of an experiment, they should instead consider applying advanced ML techniques to archival data. ML can push boundaries and motivate theoretical developments by BSci researchers in OM.

3.2.2 Design Experiment. ML techniques can expand experiment types and improve their realism by allowing scholars to analyze new forms of data. For example, text analysis allows for rigorous and scalable analysis of communication between lab participants. BSci experiments already have this functionality, but to our knowledge, the use of text analysis in such designs has not been fully leveraged. For instance, Leider and Lovejoy (2016) consider chat messages between participants in a three-tier supply chain as they bargain over prices. Despite this innovative experimental feature, the authors randomly select 60 chat

samples for qualitative analysis. Experiments related to other OM problems, such as joint decision-making (Li et al. 2019), could allow participant conversations or the use of "voice to text" software to generate text that can be mined using advanced ML techniques.

In settings where one may want to use available text (such as product descriptions or customer interactions) as stimuli, making valid inferences controlling for nuisance variables often requires simplifying the text and ensuring limited variation in text stimuli across conditions. Mukherjee et al. (2023) enable the use of many real-world, textual product descriptions as stimuli without simplification by combining deep learning and an LLM (in a tool called "Lab-GPT") to generate interpretable low-dimensional, numerical representations of unstructured text, which can then be used for experimental design. Their experiment includes showing 1,000 consumers 50,000 wine descriptions randomly sampled from the almost 120,000 in the market to understand the causal influence of a product description's focus on consumer responses.

BSci researchers could measure emotion – for example, in response to OM-relevant scenarios – using affective computing, possibly combined with other technologies, such as skin conductance and eye tracking. Some relevant scenarios where emotion may pay a role include Allon and Hanany (2012), who use a game theoretic model to derive conditions under which cutting-in-line emerges in equilibrium, Kremer and Debo (2016), who explore queue length as an indicator of quality, and Buell (2021), who studies last-place aversion. These techniques also facilitate physical experiments. For example, photo and video analysis of live interactions can be used to study natural face-to-face negotiations; trust can be studied through people's responses to discovering they were lied to (Özer et al. 2011) or that, as consumers, they paid more than someone else. Affective computing and facial recognition may also be useful to address the experimenter effect, the effect of lab participants potentially exhibiting higher levels of trust and trustworthiness because they know they are being watched despite the anonymizing of their responses. Facial recognition could catch instances in which their face displays a negative emotion that contradicts their recorded decision. Complementing data from decisions, for instance, in "trust" settings with ML techniques, may improve the efficacy of BSci experiments. Last, ML techniques can be used with observational data, such as security footage (e.g., Lu et al. 2013, Wang and Zhou 2018), for behavioral insights when traditional experiments are not feasible.

3.2.3 Analyze Data. In analyzing data, whether experimental or observational, ML techniques belong in the toolkit of BSci researchers, although they generally solve a different problem than traditional methods. Econometrics and statistics typically emphasize how a change in a covariate affects the dependent variable through, for example, a regression coefficient. With ML, the goal is often to predict

the dependent variable using the covariates. The benefit of focusing on prediction is that ML automatically discovers non-linear relationships among variables without making explicit assumptions about the underlying data generation. By contrast, in regression modeling, the researcher must usually specify possible non-linearity via interaction terms. BSci researchers should consider ML techniques for tasks that can be cast as prediction problems. For example, characterizing images (e.g., "do these facial cues show the participant is angry?"), text ("is this online review discussing wait time negatively?"), or cursor movements ("is this product option and price desirable?") as in Fisher (2023) are fundamentally matters of prediction. ML techniques offer a first step of preprocessing or categorizing such data before a second stage of analysis using more traditional statistical methods.

Given the flexibility of ML techniques, it is natural to wonder how they avoid overfitting. The typical solution is a parameter that limits the model's complexity, as has been done successfully for decades in econometrics. For example, the Hodrick–Prescott (1997) filter, in which the amount of smoothing is set by a regularization parameter, is used extensively in macroeconomics to remove cycles and other short-term fluctuations from time-series. The depth of classification trees can also function as a regularization parameter. If it is set too low and the tree grows too shallow, the model becomes overly simplistic, and only the most global patterns are detectable. If the parameter is set too high and the tree is overgrown, the estimated model becomes sensitive to noise and overfits. Only when the regularization is set appropriately, do ML methods produce accurate out-of-sample predictions.

Assessing whether an ML model produces accurate out-of-sample predictions requires accessing multiple independent data sets, which is often difficult. The standard way to determine appropriate regularization levels is thus through cross-validation: a random subset of the data is used to estimate the model, with the remaining data used to assess accuracy. This process is repeated many times, and the best-performing regularization parameter is selected. Subject to mild assumptions, cross-validation is an asymptotically optimal way of tuning and selecting models (Arlot and Celisse 2009). There are no finite-sample results, placing intuition and heuristics at the heart of many ML design decisions.

When simultaneously modeling different data types, for example, text or images along with traditional covariates, the number of variables can grow quickly, creating an estimation challenge. The number of parameters in a linear model is often greater than the number of observations when working with unstructured data or multiple sensors, such as for skin conductance or eye tracking. A popular solution for variable selection in OM (Ang et al. 2016, Ryzhov et al. 2016, Li et al. 2018, Camerer et al. 2019) uses regularization with regression via the adaptive least absolute shrinkage and selection operator

(LASSO; Zhao and Yu 2006), which beyond solving the dimensionality issue, provides consistent variable selection (Zhao and Yu 2006) and unbiased point estimates for relatively large effect sizes.

Because ML often concerns prediction, the lack of rigorous statistical inference of treatment effects (confidence intervals, hypothesis testing) can represent a limitation for BSci researchers, though developments in trustworthy and interpretable ML systems are promising. Suppose a deep learning model can predict human decision-making well; to understand *why* the algorithm and, arguably, by extension the human, made a particular prediction, methods like LIME (Ribeiro et al. 2016) and Shapley values (Lundberg and Lee 2017) can be used to assign reason-codes (e.g., "because the subject was male and exposed to treatment X") to otherwise black-box predictions. These methods support a particular behavioral mechanism but fall short of supporting formal statistical inference. ML methods are also being incorporated into traditional econometrics aimed at inference. The first stage of the instrumental variables approach is essentially a prediction step for which ML can help establish strong instruments (Li et al. 2018). Causal forests (Athey et al. 2019), a recent variant of random forests for matching treatment and control observations, can help improve statistical efficiency and power when estimating heterogeneous treatment effects.

Lastly, there are innovative applications of ML to support research replication and reproducibility. For example, *Science* recently announced its journals will begin using a commercial software called ProofFig that automates the process of detecting improperly manipulated images (Thorp 2024). Such tools could be built or adopted in OM to identify data irregularities or other problems before publication. As regards the much-cited replication crisis in the social sciences (to be fair, there are questions around replication in ML as well, see Hutson (2018)), ML could potentially be used to identify results most in need of replication using a range of criteria; for example, especially counterintuitive results or those that are widely cited could be prioritized. As Davis et al. (2023) show, people are not very successful in predicting replication outcomes, and ML may be better at identifying potentially fragile results.

4. Illustrating How ML and BSci Can Enhance Each Other in OM: Examples

In our discussion thus far, we have outlined the potential complementarities of ML and BSci. Here, we review studies successfully employing their methods or results in combination to yield valuable OM insights. We include a range of OM applications that demonstrate the flexibility of the approach: order fulfillment, physical queues, forecasting, data product development, worker performance, and food services. We offer these as single concrete examples rather than as a comprehensive literature review.

Order Fulfillment: In many modern fulfillment environments, an algorithm prescribes how workers should pack items, often to optimize volume utilization. Workers often deviate from these instructions. Using data from the Alibaba Group, Sun et al. (2022) find that packing workers do so 5.8% of the time, leading to longer packing times and lower efficiency. They posit two hypotheses for these deviations. One is that workers have more information about an order and, therefore, offer different solutions. The other, based on BSci findings, is that human decision-makers cognitive limitations and bounded rationality lead workers to respond to complex instructions from the algorithm by opting for a simpler approach (also akin to algorithm aversion), in this case, switching to a larger box to more easily fit the items.

Sun et al. (2022) propose a "human-centric bin-packing algorithm" from ML that incorporates potential worker deviations and captures 43 product and order features tied to worker deviations and box switching. For these "targeted packages," the algorithm adjusts its original instructions by lowering the maximum fill rate. Using methods from BSci, the authors implement a large-scale field experiment in conjunction with Alibaba and find that their algorithm produces a lower rate of box switching (29.5% to 23.8%) and a lower mean packing time (4.5%).

Physical Queues: Lu et al. (2013) examine the effect of waiting in physical queues on customer purchases. As discussed in Section 3.2, BSci methods can be used to study such "waiting" problems theoretically, but empirical validation is challenging, especially when experiments are involved (e.g., waiting can be odd in a lab setting, when the duration of sessions must be planned upfront). The authors address this using ML methods, notably image recognition in video recordings of a deli and a customer transaction data set.

Among other results, this approach in Lu et al. (2013) reveals that the number of customers in a queue has a meaningful effect on purchase likelihood. Moderate increases in the queue length can reduce sales by the equivalent of a 5% price increase. More importantly, their results yield an interesting behavioral insight: compared to queue length, service capacity, which directly impacts the rate at which the queue moves, has a negligible impact on sales; that is, customers rely on the visual length of the queue rather than the rate at which it moves.

Forecasting: OM researchers have begun to examine how an algorithm's prediction and a human's judgment can be combined into a forecast. Consider a healthcare setting with a hospital requiring forecasts of surgery duration for scheduling purposes. Ibrahim and Kim (2019) show an algorithmic forecast has a better absolute forecast error (in percentage terms) than those generated by physicians, 29% versus 33%; however, incorporating physicians' forecasts in the algorithm improves accuracy. This is because humans may have access to private information that is not incorporated into the algorithm.

Ibrahim et al. (2021) build on this by asking what specific type of human judgment can most improve forecasts, showing that asking for a human forecaster's private information adjustment (PIA) – how much they believe the algorithm should alter its forecast – leads to more accurate forecasts than simply asking them for their direct forecast (DF).

Ibrahim et al. (2021) arrive at their result through a combination of BSci and ML approaches. Specifically, they begin with a theoretical analysis of the algorithm using human PIA or human DF as an input. They demonstrate that the advantage of eliciting PIA over DF is larger when public data – which the algorithm can use – are complex and difficult to process but smaller when the human's private information is challenging to comprehend. This approach allows the authors to generate prescriptions without data and then administer controlled human-subject experiments in which they elicit human judgments (PIA or DF) for 50 simulated surgeries based on predictive data; participants were aware that the algorithm had access to only part of the data (public), but that they (the participant) had access to additional private data. The authors validate their theoretical predictions; the average root mean squared error was 21% lower when the algorithm used PIA versus DF.

Data Product Development: As technology improves, so does the potential for unique ways to combine ML and BSci to solve interesting OM problems. There is now a wide range of ML algorithms and systems sold as data products, including voice recognition systems, self-driving vehicles, and image recognition. Such products require new user data to be incorporated back into the algorithms used. Yet, many companies lack the expertise to successfully do this, often resorting to outsourcing from third parties. Gurkan and de Véricourt (2022) dub this the "AI flywheel effect," offering examples of the use of behavioral analytical modeling to improve ML products. They develop a two-period model that includes an accuracy versus revenue tradeoff, outsourcing incentive issues, and the impact of data on these. At the beginning of each period, the company outsources the development of an ML algorithm to a third party. The accuracy of the algorithm varies depending on the company's effort and the amount of available data. The company advertises the product to users, demand is realized, and profits are earned.

Among other insights, Gurkan and de Véricourt (2022) observe that the company's decisions largely rely on the interaction between the volume of training data and the third-party effort. If this effort has a significant impact on accuracy for large volumes of data, the company will underprice the product to obtain more data from users; it will overprice and collect less data when its effort is most useful for a lower volume of data.

Worker Performance: Many companies rely on LLMs to complete tasks, which may differ in complexity, context, and more. Dell'Acqua et al. (2023) conduct a field experiment with the Boston

Consulting Group to determine how worker performance varies with LLM assistance in different knowledge-intensive tasks (see Sections 3.1.1 and 3.2.1). Their experiment includes over 700 consultants as participants and comprises three treatments that differ with respect to access to an LLM (GPT-4): no GPT-4 access, GPT-4 access, and GPT-4 access with a prompt overview. For all treatments, they consider various tasks that are common in a consulting environment. Their results indicate that for tasks at the "frontier of AI capabilities," worker performance increases, both in terms of productivity and quality. However, for tasks outside of the frontier (i.e., where the LLM makes an error), consultants that can access the LLM are less likely to arrive at the correct solution.

Food Services: Bastani et al. (2022) study how to improve human actions in sequential decision-making environments. Recognizing that humans are necessary for certain tasks, such as food service, the authors design a novel ML algorithm that extracts best practices from data and converts those into "tips" for workers. The algorithm identifies tips that best narrow the deviation between a human's actions and those of the optimal policy while accounting for the actions that most improve overall performance. Before deploying their ML algorithm in practice, they evaluate it using human-subject experiments in which human participants assign tasks to virtual workers in a kitchen context. Several tradeoffs are involved, such as assigning a task to a worker who is available but who may be slow at that particular task versus waiting for another better-suited task. The authors manipulate whether a participant is shown any tips, tips from the algorithm, or those most suggested by humans who have already played the game.

Bastani et al.'s (2022) valuable design allows the authors to quantify changes in performance from their algorithm in a controlled environment. They can also track the rate at which the human participants follow the tips. They observe that their algorithm improves human performance through faster learning (and that human participants often combine the algorithm's tips with their own experience). ML researchers may appreciate this study's novel ML algorithm, but it is worth noting it also won the INFORMS "Best Behavioral Operations Working Paper Award" in 2021, striking an excellent balance between ML and BSci methods and results.

5. Conclusion

In this article, we discuss how the best of two fields, ML and BSci, can be applied within OM. Traditionally, these have been treated as distinct fields in OM despite overlapping objectives of prediction, causal inference, and improved OM understanding. We see the two fields as working in tandem for more rigorous and robust OM results. ML and BSci need not be exactly "equal partners" to work well together; the main impetus may come more naturally from one side or another. The 2x2 framework in Table 2

characterizes which may be more dominant based on the research objective (causal inference or prediction) and data availability (rich versus poor).

The top-left quadrant shows the situation of ample data with the objective of causal inference. Here, ML and BSci serve in supporting roles to standard statistical inference.[2] For instance, after initial results are generated by difference-in-differences (DiD) and propensity score matching, ML and BSci techniques can be used to enhance rigor and accuracy. As discussed in Section 3.2.3, causal forest techniques can improve the matching of treatment and control observations, and ML can help identify strong instruments. Controlled experiments can be used to determine the robustness of any cause-and-effect conclusions, and the BSci literature can elucidate underlying mechanisms.

In the top-right quadrant of Table 2, the situation in which there is sufficient data and the goal is prediction, ML methods are the primary methodology. BSci can assist in obtaining testing data for out-of-sample applications, evaluating the algorithm's performance, and studying deployment (particularly when interacting with a human decision-maker).

Table 2. A Classification Framework for Causal Inference and Prediction Based on Data Availability

	Causal Inference	**Prediction**
Data rich (On phenomena of interest)	ML and BSci both supportive (E.g., ML models for facilitating causal inference, BSci experiments for mechanism)	ML primary, BSci supportive (E.g., ML for algorithm design, BSci for testing data and performance assessment)
Data poor (On phenomena of interest)	BSci primary, ML supportive (E.g., BSci experiments for effect, ML techniques for experimental design)	ML and BSci both primary (E.g., ML models with synthetic or related data, BSci experiments for mechanism)

In the lower-left quadrant, where the goal is causal inference, but data are scarce, BSci plays a primary role, with support from ML. A set of controlled experiments can generate data to identify cause and effect, but ML methods can be useful in experimental design. For instance, as discussed in Section 3.2.2, ML text analysis or affective computing can be used to test hypotheses or identify factors to manipulate among experimental treatments. The lower right quadrant, seeking prediction with little data, is a challenging environment, and ML and BSci can be valuable in concert. ML models can be prototyped or initially trained with data generated from lab experiments, and once deployed, the underlying algorithms could be refined over time as the setting transitions to data rich (e.g., see the case of Blue River in Section 3.1.1).

How can we encourage such synergies in practice? A key issue is Ph.D. student training and ensuring future researchers are adequately prepared to employ these various methods. It would be unreasonable to take a full load of courses across ML and BSci. Currently, OM students take the required courses and then drift into their siloed specialties, limiting potential collaboration; ML students take advanced courses in statistics and computer science, and BSci students are immersed in economics and psychology.

Requiring ML students to take one or two behavioral courses and vice versa would result in a better balance; a dedicated course integrating the two would be ideal. One could even imagine future OM researchers becoming experts in fields that combine both, such as "behavioral ML" or "behavioral analytics."

Researchers can draw on existing resources and conferences to learn about each field. ML researchers can join BSci societies, such as the Behavioral Operations Management section of *INFORMS* and the College of Behavior In Operations Management of the Productions and Operations Management Society, attend the annual Behavioral Operations Conference, or acquire useful "one-stop" sources, such as *The Handbook of Behavioral Operations* (Donohue et al. 2019). By the same token, BSci researchers could become members of the ML chapters of INFORMS and *POM* and attend symposia and events on ML, such as the INFORMS Data Science Workshop and the Data Mining Symposium hosted by the INFORMS College on AI and its Data Mining Society, respectively. Scholars can follow developments in core ML outlets associated with, for example, the Association for the Advancement of Artificial Intelligence or the Association for Computing Machinery. We recognize that allocating their scarce time across disciplines may be challenging for researchers, but a working knowledge of the other's field will allow ML/BSci collaborations that comprehensively address important OM problems.

Acknowledgements

We are grateful to Sendhil Mullainathan, who presented a seminar on "Why Behavioral Science Needs AI and Why AI Needs Behavioral Science" at UCLA on December 8, 2023, while we were working on this revision, for pointing us to several important papers in this domain. We also acknowledge helpful comments on our article from Kris Ferreira.

Appendix

Appendix A: BSci Resources for ML Researchers

There are several BSci resources for interested ML researchers. For those wishing to run human-subject experiments but requiring assistance, there are third-party platforms that provide a suite of services; for example, SoPHIELabs offers experimental design, programming, and data collection services.[3]

Researchers wishing to design, program, and/or administer experiments can access the many resources on experimental design; a useful reference is *The Handbook of Behavioral Operations Management* (Donohue et al. 2019). Following experiment design (e.g., number of treatments, focus variables, and levels), the researcher must program the experimental game; oTree is overtaking zTree in popularity, as it is web-based and relies on a Python code format, making it especially accessible.[4] The

final step is to administer the experiment and collect data. Two examples of third-party platforms that can help in this regard are Amazon MTurk and Prolific.[5] Another option for data collection is university labs. In particular, most business schools, economics departments, or psychology departments have a dedicated laboratory that is available for researchers to run experiments.

For ML scholars not yet ready to conduct a research experiment there are predesigned games online, including on websites accessible by academic faculty and typically used for teaching purposes (classroom exercises with students as participants). For instance, Charles Holt developed "Veconlab: Experimental Economics Laboratory," which includes games that are ready to play through a web interface, including the ultimatum game and the newsvendor task.[6]

Appendix B: ML Resources for BSci Researchers

There are myriad open-source ML resources for BSci researchers, and we provide a starting point in the form of commonly used libraries and frameworks. Given that constantly evolving software landscape, we suggest researchers working in R use the task-based categories on CRAN, the network of mirrors for R, to identify relevant software libraries; for example, the curated lists of open-source R libraries include general ML methods, including to support causal inferences and text analysis.[7] Other useful outlets are *The R Journal* and the *Journal of Statistical Software*, peer-reviewed journals focusing on the implementation of statistical and ML models, including, for example, latent Dirichlet allocation (Hornik and Bettina 2011) and structural topic modeling (Roberts et al. 2016, 2019). Such curated lists, to the best of our knowledge, do not exist for Python, although there are several popular libraries we recommend. Specifically, genism, nltk, and vaderSentiment are open-source libraries containing estimation routines for topic modeling, sentiment analysis, and other text analysis and for tasks such as data preprocessing.[8]

Python is currently the most popular language for image and video analysis and most modern text analysis, like BERT and other large language models.[9] Google's well-documented and popular open-source Python framework for deep neural networks, Tensorflow, is the basis for many algorithms and the Open Computer Vision Library, or OpenCV, offers a library of functions for image and video analysis that can be useful for object and gesture detection, emotion analysis, facial recognition, and so forth.[10]

---

[1] Note that our definition of external validity does not speak to the replicability of results in subsequent studies. We discuss replication in more detail in Section 3.2.

[2] We recognize that other methodologies may be useful in these settings but, for reasons of brevity, these have been excluded from the scope of our study.

[3] https://www.sophielabs.com/

[4] https://www.ztree.uzh.ch/en.html and https://www.otree.org/

[5] https://www.mturk.com/ and https://www.prolific.com/
[6] https://veconlab.econ.virginia.edu/
[7] See https://cran.r-project.org/web/views/; https://cran.r-project.org/web/views/NaturalLanguageProcessing.html; https://cran.r-project.org/web/views/MachineLearning.html.
[8] See https://radimrehurek.com/gensim/; https://www.nltk.org/; https://github.com/cjhutto/vaderSentiment.
[9] See https://www.tensorflow.org/text/tutorials/classify_text_with_bert.
[10] See https://www.tensorflow.org/; https://opencv.org/.

References

Agrawal, A., J. Gans, and A. Goldfarb (2018), Prediction Machines: The Simple Economics of Artificial Intelligence. Brighton, MA: Harvard Business Review Press.

Allon, Gad, Eran Hanany. 2012. "Cutting in Line: Social Norms in Queues." *Management Science* 58(3), 493-506.

Agan, Amanda Y., Diag Davenport, Jens Ludwig, and Sendhil Mullainathan. 2023. Automating automaticity: How the context of human choice affects the extent of algorithmic bias. National Bureau of Economic Research, No. w30981.

Ang, Erjie, Sara Kwasnick, Mohsen Bayati, Erica L. Plambeck, Michael Aratow. 2016. "Accurate Emergency Department Wait Time Prediction." *Manufacturing & Service Operations Management*, 18(1), 141-156.

Arlot, Sylvain, Alain Celisse. 2010. "A Survey of Cross Validation Procedures for Model Selection." *Statistics Surveys*, 4, 40-79.

Athey, Susan, Julie Tibshirani, Stefan Wager. 2019. "Generalized Random Forests." *The Annals of Statistics* 47(2), 1148-1178.

Babic, Boris, I. Glenn Cohen, Theodoros Evgeniou, Sara Gerke. 2021. "When Machine Learning Goes Off the Rails." *Harvard Business Review*, https://hbr.org/2021/01/when-machine-learning-goes-off-the-rails.

Bai, Bing, Hengchen Dai, Dennis J. Zhang, Fuqiang Zhang, Haoyuan Hu. 2022. "The Impacts of Algorithmic Work Assignment on Fairness Perceptions and Productivity: Evidence from Field Experiments." Working paper.

Bastani, Hamsa, Osbert Bastani, Wichinpong Park Sinchaisri. 2022. "Improving Decision-Making with Machine Learning." Working paper.

Bearden, William O., Richard G. Netemeyer. 1999. "Handbook of Marketing Scales: Multi-item Measures for Marketing and Consumer Behavior Research." Sage Publications Inc., California, USA.

Blei, David M., Andrew Y. Ng, Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, 3, 993-1022.

Blythe, Stephen. 2018. "Big Data and Machine Learning Won't Save Us from Another Financial Crisis." *Harvard Business Review*, https://hbr.org/2018/09/big-data-and-machine-learning-wont-save-us-from-another-financial-crisis.

Bolton, Gary E., Yefen Chen. 2019. "Other-regarding Behavior: Fairness, Reciprocity, and Trust." in *The Handbook of Behavioral Operations*, edited by Karen Donohue, Elena Katok, and Stephen Leider, John Wiley & Sons, Hoboken, NJ.

Bolton, Gary E., Elena Katok. 2018. "Cry Wolf or Equivocate? Credible Forecast Guidance in a Cost-Loss Game." *Management Science*, 64(3), 1440-1457.

Bolton, Gary E., Axel Ockenfels, Ulrich W. Thonemann. 2012. "Managers and Students as Newsvendors." *Management Science*, 58(12), 2225-2233.

Boyaci, Tamer, Caner Canyakmaz, Francis de Véricourt. 2023. "Human and Machine: The Impact of Machine Input on Decision Making Under Cognitive Limitations." *Management Science*, forthcoming.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. 2020. "Language models are few-shot learners." *Advances in neural information processing systems,* 33: 1877-1901.

Buell, Ryan W. 2018. "The Parts of Customer Service That Should Never Be Automated." *Harvard Business Review*, https://hbr.org/2018/02/the-parts-of-customer-service-that-should-never-be-automated

Buell, Ryan W. 2021. "Last-place Aversion in Queues." *Management Science*, 67(3), 1430-1452.

Buell, Ryan W., Tami Kim, Chia-Jung Tsay. 2017. "Creating Reciprocal Value Through Operational Transparency." *Management Science*, 63(6), 1673-1695.

Burton, Jason W., Mari-Klara Stein, Tina Blegind Jensen. 2020. “A Systematic Review of Algorithm Aversion in Augmented Decision Making.” *Journal of Behavioral Decision Making*, 33, 220-239.

Camerer, Colin F. 2003. “Behavioral Game Theory: Experiments in Strategic Interaction.” Princeton University Press, Princeton NJ.

Camerer, Colin F., Gideon Nave, and Alec Smith. 2019. "Dynamic unstructured bargaining with private information: theory, experiment, and outcome prediction via machine learning." *Management Science,* 65(4): 1867-1890.

Caro, Felipe, Anna Saez de Tejada Cuenca. 2018. “Believing in Analytics: Managers’ Adherence to Price Recommendations from a DSS.” Working paper.

Chen, Ningyuan, Ming Hu, Wenhao Li. 2022. “Algorithmic Decision-Making Safeguarded by Human Knowledge.” arXiv:2211.11028.

Corbett, Charles J., Luk N. Van Wassenhove. 1993. “The Natural Drift: What Happened to Operations Research?” *Operations Research*, 41(4), 625-640.

Cukier, Kenneth, Viktor Mayer-Schönberger, Francis de Véricourt. 2021. “Framers: Human Advantage in an Age of Technology and Turmoil.” Penguin Random House, NY.

Cui, Yao, Andrew M. Davis. 2022. “Tax-Induced Inequalities in the Sharing Economy.” *Management Science*, 68(10), 7202-7220.

Cui, Tony Haitao, Yaozhong Wu. 2019. “Incorporating Behavioral Factors into Operations Theory.” in *The Handbook of Behavioral Operations*, edited by Karen Donohue, Elena Katok, and Stephen Leider, John Wiley & Sons, Hoboken, NJ.

Dai, Tinglong, Shubhranshu Singh. 2023. “Artificial Intelligence on Call: The Physician’s Decision of Whether to Use AI in Clinical Practice.” Working paper.

Davis, Andrew M. 2019. “Biases in Individual Decision-Making.” in *The Handbook of Behavioral Operations*, edited by Karen Donohue, Elena Katok, and Stephen Leider, John Wiley & Sons, Hoboken, NJ.

Davis, Andrew M., Blair Flicker, Kyle Hyndman, Elena Katok, Samantha Keppler, Stephen Leider, Xiaoyang Long, Jordan D. Tong. 2023. “A Replication Study of Operations Management Experiments in *Management Science*.” *Management Science*, 69(9), 4977-4991.

de Véricourt, Francis, Huseyin Gurkan. 2023. “Is Your Machine Better Than You? You May Never Know.” *Management Science*, forthcoming.

de Véricourt, Francis, Georgia Perakis. 2020. “Frontiers in Service Science: The Management of Data Analytics Services: New Challenges and Future Directions.” *Service Science*, 12(4), 121-129.

De-Arteaga, Maria, Stefan Feuerriegel, and Maytal Saar-Tsechansky. 2022. “Algorithmic Fairness in Business Analytics: Directions for Research and Practice.” *Production and Operations Management*, 31(10), 3749-3770.

Dell’Acqua, Fabrizio, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, Francois Candelon, Karim R. Lakhani. 2023. “Navigating the Jagged Technological Frontiers: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.” *Working Paper No. 24-013*, Harvard Business School.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding." In Proceedings of NAACL-HLT, 4171-4186.

Dietvorst, Berkeley J., Joseph P. Simmons, Cade Massey. 2015. “Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err.” *Journal of Experimental Psychology: General* 144(1), 114-126.

Donohue, Karen, Elena Katok, Stephen Leider. 2019. “The Handbook of Behavioral Operations.” Edited by Karen Donohue, Elena Katok, and Stephen Leider, John Wiley & Sons, Hoboken, NJ.

Eberhardt, Frederick, Nur Kaynar, Auyon Siddiq. 2023. “Discovering Causal Models with Optimization: Confounders, Cycles, and Feature Selection.” *Management Science*.

Fisher, Geoffrey. 2023. "Measuring the factors influencing purchasing decisions: Evidence from cursor tracking and cognitive modeling." *Management Science*.

Gavirneni, Srinagesh, Alice M. Isen. 2010. “Anatomy of a Newsvendor Decision: Observations from a Verbal Protocol Analysis.” *Production and Operations Management*, 19(4), 453-462.

Goodell, John W., Satish Kumar, Weng Marc Lim, Debidutta Pattnaik. 2021. “Artificial Intelligence and Machine Learning in Finance: Identifying Foundations, Themes, and Research Clusters from Bibliometric Analysis.” *Journal of Behavioral and Experimental Finance*, 32, 100577.

Gurkan, Huseyin, Francis de Véricourt. 2022. “Contracting, Pricing, and Data Collection Under the AI Flywheel Effect.” *Management Science*, 68(12), 8791-8808.

Hodrick, Robert J., Edward C. Prescott. 1997. “Postwar U.S. Business Cycles: An Empirical Investigation.” *Journal of Money, Credit, and Banking*, 29(1), 1-16.

Hornik, Kurt, and Bettina Grün. 2011. "topicmodels: An R Package for Fitting Topic Models." *Journal of Statistical Software*, 40(13), 1-30.

Horton, John J. 2017. “The Effects of Algorithmic Labor Market Recommendations: Evidence from a Field Experiment.” *Journal of Labor Economics*, 35(2), 345-385.

Hunt, Davis. 2021. “Can Algorithms be Ethical?” *ORMS Today*, https://doi.org/10.1287/orms.2021.06.12

Hutson, Matthew. 2018. “Artificial Intelligence Faces Reproducibility Crisis.” *Science*, 359(6377) 725-726.

Ibanez, Maria R., Bradley R. Staats. 2019. “Behavioral Empirics and Field Experiments.” in *The Handbook of Behavioral Operations*, edited by Karen Donohue, Elena Katok, and Stephen Leider, John Wiley & Sons, Hoboken, NJ.

Ibrahim, Rouba, Song-Hee Kim. 2019. “Is Expert Input Valuable? The Case of Predicting Surgery Duration.” *Seoul Journal of Business*, 25(2), 1-34.

Ibrahim, Rouba, Song-Hee Kim, Jordan Tong. 2021. “Eliciting Human Judgment for Prediction Algorithms.” *Management Science*, 67(4), 2314-2325.

Johnsen, Lennart C., Guido Voigt, Charles J. Corbett. 2019. “Behavioral Contract Design Under Asymmetric Forecast Information.” *Decision Sciences*, 50(4), 786-815.

Kagel, John H., Alvin E. Roth. 2016. “The Handbook of Experimental Economics, Volume 2.” Princeton University Press, Princeton NJ.

Kahneman, Daniel. 2011. “Thinking Fast and Slow.” Farrar, Straus, and Giroux, New York NY.

Kahneman, Daniel, Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision Under Risk.” *Econometrica*, 47(2), 263-292.

Kelliher, Rebecca. 2021. “AI in Admissions Can Reduce or Reinforce Biases.” *Diverse, Issues in Higher Education*, https://www.diverseeducation.com/students/article/15114427/ai-in-admissions-can-reduce-or-reinforce-biases.

Kesavan, Saravanan, Tarun Kushwaha. 2020. “Field Experiment on the Profit Implications of Merchants’ Discretionary Power to Override Data-Driven Decision-Making Tools.” *Management Science* 66(11), 5182-5190.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan. 2018. “Human Decisions and Machine Predictions.” *The Quarterly Journal of Economics*, 133(1), 237-293.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Manish Raghavan. 2023a. The inversion problem: Why algorithms should infer mental state and not just predict behavior. *Perspectives on Psychological Science,* 17456916231212138.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2023b. The challenge of understanding what users want: Inconsistent preferences and engagement optimization. *Management Science*.

Klockmann Victor, Alicia von Schenk, Marie Claire Villeval. 2022. "Artificial Intelligence, Ethics, and Intergenerational Responsibility." *Journal of Economic Behavior & Organization*, 203, 284-317.

Kremer, Mirko, and Laurens Debo. 2016. "Inferring Quality from Wait Time." *Management Science* 62(10), 3023-3038.

Larson, Erik J. 2021. "The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do." *Harvard University Press*, 2021.

LeCun, Yann, Yoshua Bengio, Geoffrey Hinton. 2015. "Deep learning." *Nature,* 521.7553: 436-444.

Lee, Yun Shin, Enno Siemsen. 2017. "Task Decomposition and Newsvendor Decision Making." *Management Science*, 63(10), 3226-3245.

Leider, Stephen, William S. Lovejoy. 2016. "Bargaining in Supply Chains." *Management Science* 62(10), 3039-3058.

Li, Jiawei, Damian R. Beil, Stephen Leider. 2019. "Team Decision Making in Operations Management." Working paper.

Li, Jun, Serguei Netessine, Sergei Koulayev. 2018. "Price to Compete… with Many: How to Identify Price Competition in High-Dimensional space." *Management Science*, 64(9), 4118-4136.

Libby, Robert, Robert Bloomfield, Mark W. Nelson. 2002. "Experimental Research in Financial Accounting." *Accounting, Organizations and Society*, 27(8), 775-810.

Liu, Bing. 2020. "Sentiment Analysis: Mining Opinions, Sentiments, and Emotions." Cambridge University Press.

Loughran, Tim, Bill McDonald. 2011. "When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance* 66(1), 35-65.

Lu, Yina, Andres Musalem, Marcelo Olivares, Ariel Schilkrut. 2013. "Measuring the Effect of Queues on Customer Purchases." *Management Science*, 59(8), 1743-1763.

Luca, Michael, John Kleinberg, Sendhil Mullainathan. 2016. "Algorithms Need Managers, Too." *Harvard Business Review*, https://hbr.org/2016/01/algorithms-need-managers-too

Lundberg, Scott M., Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems*, 30.

Mejia, Jorge, Shawn Mankad, Anandasivam Gopal. 2021. "Service Quality Using Text Mining: Measurement and Consequences." *Manufacturing & Service Operations Management* 23(6), 1354-1372.

Mišić, Velibor V., Georgia Perakis. 2020. "Data Analytics in Operations Management: A Review." *Manufacturing & Service Operations Management*, 22(1), 158-169.

Mukherjee, Anirban, Hannan Chang, and Sachin Gupta. 2023. "Bridging the Gap: Using Interpretable AI to Incorporate Real-World Product Descriptions in Consumer Research Experiments," working paper, Cornell University.

Mullainathan, Sendhil, and Ashesh Rambachan. 2023. From Predictive Algorithms to Automatic Generation of Anomalies. Available at SSRN 4443738.

Niu, Shuteng, Yongxin Liu, Jian Wang, and Houbing Song. 2020. "A decade survey of transfer learning (2010–2020)." *IEEE Transactions on Artificial Intelligence*, 1(2): 151-166.

Ng, Andrew. 2018. AI Transformation Playbook. Accessed December 18, 2023, https://www.coursera.org/lecture/ai-for-everyone/ai-transformation-playbook-part-2-kSWz6

O'Neil, Cathy. 2016. “Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.” Crown Books.

Özer, Özalp, Yanchong Zheng, Kay-Yut Chen. 2011. "Trust in Forecast Information Sharing." *Management Science* 57(6), 1111-1137.

Pan, Sinno Jialin, and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), pp.1345-1359.

Pasquale, Frank. 2015. “The Black Box Society: The Secret Algorithms that Control Money and Information.” Harvard University Press.

Picard, Rosalind. 2000. “Affective Computing.” MIT Press, Cambridge, MA.

Rajkomar, Alvin, Jeffrey Dean, Isaac Kohane. 2019. “Machine Learning in Medicine.” *The New England Journal of Medicine*, 380(14), 1347-1358.

Ren, Yufei, Rachel Croson. 2013. “Overconfidence in Newsvendor Orders: An Experimental Study.” *Management Science*, 59(11), 2502-2517.

Ribeiro, Marco Tulio, Sameer Singh, Carlos Guestrin. 2016, August. “‘Why Should I Trust You?’ Explaining the Predictions of Any Classifier.” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.

Roberts, Margaret E., Brandon M. Stewart, Edoardo M. Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." Journal of the American Statistical Association 111(515), 988-1003.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. "stm: An R package for Structural Topic Models." Journal of Statistical Software, 91(2), 1-40.

Roels, Guillaume, Bradley R. Staats. 2021. “OM Forum—People-Centric Operations: Achievements and Future Research Directions.” *Manufacturing & Service Operations Management*, 23(4), 745-757.

Ryzhov, Ilya O., Bin Han, Jelena Bradić. 2016. “Cultivating Disaster Donors Using Data Analytics.” *Management Science*, 62(3), 849-866.

Settles, Burr. 2009. “Active Learning Literature Survey.” Computer Sciences Technical Report 1648 University of Wisconsin–Madison

Siemsen, Enno, John Aloysius. 2020. “Supply Chain Analytics and the Evolving Work of Supply Chain Managers.” White paper for the *Association for Supply Chain Management*.

Soll, Jack B., Katherine L. Milkman, John W. Payne. 2015. “A User’s Guide to Debiasing.” in *The Wiley Blackwell Handbook of Judgment and Decision Making*, edited by Gideon Keren and George Wu, John Wiley & Sons.

Sun, Jiankun, Dennis J. Zhang, Haoyuan Hu, Jan A. Van Mieghem. 2022. “Predicting Human Discretion to Adjust Algorithmic Prescription: A Large-Scale Field Experiment in Warehouse Operations.” *Management Science*, 68(2), 846-865.

Taylor, Jessica. 2016. “Quantilizers: A Safer Alternative to Maximizers for Limited Optimization.” *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence*, Technical Report WS-16-02.

Thaler, Richard H., Cass R. Sunstein. 2009. “Nudge.” Penguin Books, New York NY.

Thorp, Holden H. 2024. “Genuine images in 2024.” *Science* 383, 7-7. DOI:10.1126/science.adn7530.

Tobe, Frank. 2017. “Blue River Technology Sells to Deere for $305 Million.” *The Robot Report*. Accessed December 18, 2023. https://www.therobotreport.com/startup-blue-river-technology-sells-deere-305-million/

Trautman, Erik. 2018. “The Virtuous Cycle of AI Products.” Accessed December 18, 2023. https://www.eriktrautman.com/posts/the-virtuous-cycle-of-ai-products

Uzialko, Adam. 2022. “Artificial Insurance? How Machine Learning is Transforming Underwriting.” *Business News Daily*, https://www.businessnewsdaily.com/10203-artificial-intelligence-insurance-industry.html.

Wang, Jingqi, Yong-Pin Zhou. 2018. “Impact of Queue Configuration on Service Time: Evidence from a Supermarket.” *Management Science*, 64(7), 3055-3075.

Weiss, Karl, Taghi M. Khoshgoftaar, and DingDing Wang. 2016. "A survey of transfer learning." *Journal of Big data*, 3, no.1: 1-40.

West, Darrell M. 2021. “Using AI and Machine Learning to Reduce Government Fraud.” *Brookings*, https://www.brookings.edu/research/using-ai-and-machine-learning-to-reduce-government-fraud/.

Zhang, Dennis, Xinyu Cao. 2022. “The Impact of Forced Intervention on AI Adoption.” Working paper.

Zhang, Lei, Shuai Wang, Bing Liu. 2018. "Deep Learning for Sentiment Analysis: A Survey." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1253.

Zhao, Peng, Bin Yu. 2006. “On Model Selection Consistency of Lasso.” *The Journal of Machine Learning Research*, 7, 2541-2563.