Received: 10 September 2018 | Revised: 9 February 2020 | Accepted: 14 February 2020 | Published on: 7 April 2020

DOI: 10.1002/smj.3152

RESEARCH ARTICLE

STRATEGIC MANAGEMENT JOURNAL WILEY

Machine learning and human capital complementarities: Experimental evidence on bias mitigation

Prithwiraj Choudhury[1] | **Evan Starr**[2] | **Rajshree Agarwal**[2]

[1]Harvard Business School, Boston, Massachusetts

[2]Robert H. Smith School of Business, University of Maryland, College Park, Maryland

Correspondence
Evan Starr, Robert H. Smith School of Business, University of Maryland, College Park, MD.
Email: estarr@rhsmith.umd.edu

Abstract

Research Summary: The use of machine learning (ML) for productivity in the knowledge economy requires considerations of important biases that may arise from ML predictions. We define a new source of bias related to incompleteness in real time inputs, which may result from strategic behavior by agents. We theorize that domain expertise of users can complement ML by mitigating this bias. Our observational and experimental analyses in the patent examination context support this conjecture. In the face of "input incompleteness," we find ML is biased toward finding prior art textually similar to focal claims and domain expertise is needed to find the most relevant prior art. We also document the importance of vintage-specific skills, and discuss the implications for artificial intelligence and strategic management of human capital.

Managerial Summary: Unleashing the productivity benefits of machine learning (ML) technologies in the future of work requires managers to pay careful attention to mitigating potential biases from its use. One such bias occurs when there is input incompleteness to the ML tool, potentially because agents strategically provide information that may benefit them. We demonstrate that in such circumstances, ML tools can make worse predictions than the prior technology vintages. To ensure productivity benefits of ML in light of potentially strategic inputs, our research suggests that

managers need to consider two attributes of human capital—domain expertise and vintage-specific skills. Domain expertise complements ML by correcting for the (strategic) incompleteness of the input to the ML tool, while vintage-specific skills ensure the ability to properly operate the technology.

KEYWORDS

bias, complementarities, domain expertise, human capital, machine learning

1 | INTRODUCTION

Artificial intelligence (AI) and machine learning (ML)—where algorithms learn from existing patterns in data to conduct statistically driven predictions and facilitate decisions (Brynjolfsson & McAfee, 2014; Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2017)—may well transform the future of work, with questions regarding whether it would substitute or complement human capital (Autor, 2015; Bughin et al., 2017; Frank et al., 2019). Despite the promise of ML in increasing productivity, many firms have encountered significant challenges due to biases in predictions,[1] often thought to result from biased training data and/or algorithms (Baer & Kamalnath, 2017; Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016; Polonski, 2018).[2]

In many important contexts, however, a third source of bias may arise because agents strategically alter the input to the algorithm, perhaps because they stand to benefit from biased predictions. For example, ML algorithms can speed up the reviewing of resumes in recruiting, or processing of insurance claims. However, resumes and insurance claims are generated by applicants who have a strategic interest in positive outcomes. Can ML correct for such strategic behavior? Research in "adversarial" ML examines attempts to "trick" ML technologies (Goodfellow, Shlens, & Szegedy, 2014), and generally concludes that it is challenging to "adversarially train" the ML technology to account for every possible input. The combination of strategically generated inputs and imperfect adversarial training of ML creates biased predictions that stem from what we term *input incompleteness*. Accordingly, two important questions arise: How can firms mitigate such bias to unlock the potential of ML? And, how may human capital complement ML to do so?

We examine answers to these questions in the context of ML technology used for patent examination, a context rife with input incompleteness. Patent examiners face a time-consuming challenge of accurately determining the novelty and nonobviousness of a patent application by sifting through ever-expanding amounts of "prior art." Moreover, patent applicants are permitted by law to create hyphenated words and assign new meaning to existing words to accurately reflect novel inventions (D'hondt, 2009; Verberne, D'hondt, Oostdijk, & Koster, 2010). However, this freedom also allows patent applicants to strategically write their applications to enhance

[1]The word "bias" evokes many connotations. We use bias to reflect inaccurate predictions (e.g., Types 1 and 2 errors).
[2]Gender and racial biases resulted in Amazon discontinuing an ML-based hiring technology (Dastin, 2018), IBM and Microsoft coming under fire for ML-based facial recognition (Buolamwini & Gebru, 2018), and the Apple credit card scrutinized through a regulatory investigation (Knight, 2019).

the likelihood of being *perceived* as novel and nonobvious by the examiner. Patent applicants can do so by including irrelevant information, or by omitting relevant citations (Alcacer & Gittelman, 2006; Lampe, 2012). Within this context, to improve the expediency and accuracy of the adjudication process, the U.S. Patent & Trademark Office (USPTO) has invested in a new ML technology (Krishna et al., 2016), which "reads" the text of a patent application to identify the most relevant prior art. Ideally, the ML tool would quickly and efficiently identify the most relevant prior art, leading to faster and more accurate decisions. However, if the application is strategically altered, the ML tool may find less relevant prior art, potentially leading to an incorrect decision. Moreover, although it is theoretically feasible for ML algorithms to continually learn and correct for ways that patent applicants attempt to manipulate the algorithm, the potential for patent applicants to dynamically update their writing strategies makes it practically impossible to adversarially train an ML algorithm to correct for this behavior.

In this study, we propose that individuals who possess domain expertise—the skills and knowledge accumulated through prior learning within a domain (Simon, 1991)— are complementary to ML in mitigating bias stemming from input incompleteness, because domain experts bring relevant outside information to correct for strategically altered inputs. We also posit that individuals with vintage-specific skills (Chari & Hopenhayn, 1991)—the skills and knowledge accumulated through prior familiarity of tasks with the technology—will be more productive because they will have higher absorptive capacity to handle the complexities in ML technology interfaces.

We provide both observational and experimental evidence in support of our thesis. The observational evidence bolsters two assumptions about our context. First, we show that patent language changes over time, even within a narrowly defined subclass, suggesting that it will be challenging for the ML tool to make quality predictions for every input. Second, we document that on the prior Boolean search technology, more experienced examiners bring more external knowledge to the adjudication process—a necessary precondition for our proposed complementarities between domain-expertise and ML in light of input incompleteness.

Our experimental design provides direct support of the hypothesized relationships by simulating patent adjudication based on identification of just one prior patent (the "silver-bullet") that should invalidate the claims of the focal application. We randomize the search technology (ML vs. Boolean) and domain expertise (expert advice from a senior USPTO examiner), and also stratify the groups by computer science and engineering (CS&E) knowledge to proxy for vintage-specific skills. We find that ML does what we would expect: relative to the prior vintage, it directs examiners to a narrower set of prior art that are much more textually similar to the focal patent application. However, because the silver bullet patent is *not* textually similar to the patent application, ML actually performs worse in finding the silver bullet than Boolean technology. Domain expertise improves the likelihood of finding the silver bullet for both technologies, and it has a stronger effect in redirecting those using ML away from patents that are textually similar to the focal application. Moreover, examiners with CS&E backgrounds perform better on ML technology, a differential driven by their ability to handle user interface complexities when implementing expert advice.

Our article contributes to several literatures. First, it contributes to the growing literature on bias in ML (Cowgill & Tucker, 2020; Polonski, 2018) by highlighting (strategic) input incompleteness as a source of bias. We expect our results to generalize to other settings where input incompleteness could conceivably be a source of biased predictions for ML algorithms, specifically when agents have incentives to game the ML technology and it is impossible to perfectly adversarially train the data. Second, amidst fierce debate about AI's future replacement of

humans at work (Autor, 2015; Benzell, Kotlikoff, LaGarda, & Sachs, 2015; Bessen, 2016; Brynjolfsson & McAfee, 2014), there is scant attention to how productivity gains from substituting ML for older vintages may be conditioned by complementary human capital. Our contribution is in highlighting the role of domain-specific expertise as a complement to ML, because it will continue to retain value when there is potential for input incompleteness. Third, our results on productivity differentials arising from vintage-specific skills contribute to the strategic management of innovation literature on pace of technology substitution (Adner & Kapoor, 2016; Adner & Snow, 2010; Agarwal & Prasad, 1999; Christensen, 1997). We show vintage-specific human capital is an important factor that may retard new technology diffusion in spite of their potential relative superiority (even with domain-expertise). However, we note the extent to which ML privileges vintage-specific human capital depends on the pace at which future iterations of the technology reduce the need for such specialized skills.

2 | BRIEF LITERATURE REVIEW

Artificial intelligence (AI) is defined as "the capability of a machine to imitate intelligent behavior" (Merriam-Webster, 2018). In its current ML form, AI is still in its infancy, and widespread adoption of ML technologies depends on their performance relative to existing vintages. In this sense, ML reflects age-old issues related to new technologies' effects on productivity, management practices, and changing skill demands for individuals at work (Autor, 2015; Bessen, 2016; Brynjolfsson & McAfee, 2014; Hounshell, 1985; Mokyr, 1990). However, ML technologies are different inasmuch as their potential for replacing humans in the workforce is untoward. The fact that machines *learn* implies cognitive processes, hitherto a domain of humans, may now be performed by technologies. Scholars have focused on two related issues critical to how ML interfaces with humans: comparative advantage of humans and machines in cognitive tasks, and relative cognitive biases of humans and machines. We turn to each below.

2.1 | Cognitive task structure and ML-human capital substitution versus complementarity

Endogenous models of technological change highlight that a new technology's diffusion relative to existing vintages depends not only to its "objective" superiority, but also on the technology's interaction with complementary resources and capabilities, particularly complementary human capital (Adner & Kapoor, 2016; Chari & Hopenhayn, 1991; Jovanovic, 2009). Each technology ushers with it changing composition of various tasks (Autor, Levy, & Murnane, 2003), so Chari and Hopenhayn (1991) define vintage-specific human capital as technology-specific skills related to both the evolution of tasks across technology vintages and the accumulation through learning-by-doing of task-specific human capital (Gibbons & Waldman, 2004). Building on this literature, debates regarding whether AI will complement or substitute humans rest on widely different assumptions in theoretical models (Aghion, Bergeaud, Blundell, & Griffith, 2017; Autor, 2015; Benzell et al., 2015). Nonetheless, several scholars provide reasons why, at least in the foreseeable future, AI's substitution for humans in cognitive tasks is overstated (Agrawal, Gans, & Goldfarb, 2018; Autor, 2014). Agrawal et al. (2018) state ML technologies can substitute humans for prediction tasks, but not for judgment tasks. Autor (2014) notes productivity gains will stem from complementarities between machines' comparative advantage in routine,

codifiable tasks, and humans' comparative advantage on tasks requiring tacit knowledge, flexibility, judgment, and creativity. The latter relates particularly to domain expertise, acquired through specialization and deliberate practice (Chase & Simon, 1973; Ericsson, Krampe, & Tesch-Römer, 1993; Simon, 1991), and linked to increased productivity (Becker, 1962; Castanias & Helfat, 1991).[3]

2.2 | Cognitive biases of humans and ML technologies

Even for predictions of routine, codifiable tasks, budding research acknowledges cognitive biases in both humans and ML technologies (Bolukbasi et al., 2016; Kleinberg et al., 2017; Polonski, 2018). In certain circumstances, ML technologies may be less susceptible than humans to cognitive biases. Kleinberg et al. (2017) show ML outperforms judges in decisions to withhold bail because judges may incorporate (unmeasured) perceptions of the defendant's demeanor into their decisions, even if the latter is a poor predictor of the defendant's propensity to show up for a postbail hearing. Cowgill (2017) shows ML hiring algorithms perform better for job performance because HR recruiters privilege applicants who are "traditional," among other reasons.

In other circumstances, scholars have documented evidence regarding two distinct biases in ML technologies: biases in the model/ML algorithm, and biases/sample selection in the training dataset (Cowgill & Tucker, 2020; Kong & Dietterich, 1995; Kleinberg et al., 2017; Osoba & Welser IV, 2017). Studies note racial bias in risk prediction tools and data used by judges in criminal sentencing (Angwin, Larson, Mattu, & Kirchner, 2016; Robinson & Koepke, 2016), and in Google's ML algorithms that yielded a greater percentage of ads with "arrest" for black versus white name searches (Sweeney, 2013). ML predictions are also fraught with sample selection due to "selective labels," whereby the training data cover only part of the population over which predictions need to be made (Lakkaraju, Kleinberg, Leskovec, Ludwig, & Mullainathan, 2017).[4]

However, a third source of cognitive bias—plaguing both humans and ML technologies—has received relatively less attention. *Input incompleteness*—when all relevant information required for search and prediction is not provided—results in salience bias inasmuch as decisions are made based on what is most salient, rather than what is most relevant. Input incompleteness could be construed as a variant of what computer scientists note as "adversarial machine learning,"[5] wherein strategic actors lead the machine into making a poor prediction. For example, in ML tools that try to classify photos, a small perturbation to the photo can dramatically alter how ML tools classify the photo, even if it looks nearly identical to the human eye (Goodfellow et al., 2014). Adversarial ML can also take the form of attempting to poison the training data in the first place (Baracaldo, Chen, Ludwig, Safavi, & Zhang, 2018).

[3]Scholars also document that non-expert or novices benefit through vicarious learning from domain experts (Greenwood, Agarwal, Agarwal, & Gopal, 2019; Thornton & Thompson, 2001). Empirical research substantiates increases in a new technology's productivity when workplace practices include provision of technology and task specific information (Bapna, Langer, Mehra, Gopal, & Gupta, 2013; Bartel, 1994; Black & Lynch, 2001). This is because codification and transmission of tacit knowledge increases speed of followers' learning from pioneers' learning-by-doing (Edmondson, Winslow, Bohmer, & Pisano, 2003).

[4]Kleinberg et al. (2017) highlight selective labels bias in training data for predicting probabilities for defendants post-bail hearing appearance, inasmuch as only those who awarded bail have the opportunity to appear for a post-bail hearing.

[5]For an easy introduction to adversarial machine learning, see: https://openai.com/blog/adversarial-example-research/

Most of the adversarial ML literature focuses on identifying the set of possible ways the machine could be misled, and then feeding the ML technology "adversarial examples" as a way to train the algorithm (Papernot et al., 2017). Ultimately, however, adversarial examples can be hard to defend against because they require ML models to produce good outputs *for every possible input*. Within this literature, one key to combatting adversarial examples is to ensure opacity of the ML algorithm, so that it is harder to game (Cowgill & Tucker, 2020). However, such solutions may reduce effectiveness and efficiencies, to the extent that transparency of ML technologies is a desirable feature for widespread diffusion. To the best of our knowledge, if and how human capital may serve as a complement to ML as a potential solution to bias arising from input incompleteness has not been examined in the social sciences and strategic management of technology literature.

3 | RESEARCH CONTEXT: U.S. PATENT EXAMINATION

The U.S. Patent and Trademark Office (USPTO) examination process is a setting fraught with input incompleteness and current search technology challenges, as we describe below.

To meet the legal standards for a patent, claims must be novel (i.e., a new, unique invention) and nonobvious (i.e., sufficiently inventive).[6] USPTO examiners review patent applications, compare it to prior art (i.e., all existing patents, patent applications, and published materials), and determine whether an inventor's claims meet these two criteria. Databases such as Patent Full-Text and Image Database (PatFT), Patent Applications Full Text (AppFT), published articles, and trade journals provide the repository of prior art patents and their claims. Prediction about novelty of claims requires the examiner to devise a search strategy and use available technologies to identify a list of potentially relevant prior art for the specific patent application (see Appendix B). The examiner then reviews the results, revises the search strategy, and ultimately decides whether to invalidate the claims and, if a patent is granted, the scope of the claims.

In maintaining a balance between expediency and patent examination quality, the USPTO aims to provide a first office action within 10 months of receipt. However, increases in the number of patent applications received by the USPTO (e.g., an 18% increase between 2010 and 2014) have created significant delays with over half a million backlogged applications, and current processing of first office action at more than sixteen months.[7] This is in spite of USPTO investing significant resources to increase the number of examiners by 28% from 2010 to 2014 (Choudhury, Khanna, & Mehta, 2017; Crouch, 2014). The Government Accountability Office report on intellectual property (GAO, 2016) notes these delays are compounded by recent increases in disputes over patent validity, raising concerns "the USPTO examiners do not always identify the most relevant prior art, which has resulted in granting some patents that may not meet the statutory requirements for patentable inventions" (p. 2). The concerns regarding time and accuracy of patent examination can be traced to the following two challenges in the various important and laborious steps in the process (GAO, 2016).

[6]For more details, please refer to Appendix B, Exhibit B1&B2, GAO (2016) and www.uspto.gov.

[7]USPTO, "U.S. Patent Statistics Chart Calendar Years 1963–2015," June 15, 2016 (accessed July 2017) https://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm,; "First Office Action Pendency," June 2017 (accessed July 2017); https://www.uspto.gov/corda/dashboards/patents/main.dashxml?CTNAVID=1004; "Patents Data, At a Glance," June 2017 (accessed July 2017), https://www.uspto.gov/dashboards/patents/main.dashxml

3.1 | Input incompleteness in the patent examination process

Because novel and nonobvious inventions may often require new language and descriptors, patent law gives applicants extensive agency in what to include or exclude from the patent application when making claims, including the ability to be their own "lexicographers." The result of this freedom is that authors of patent claims often introduce complex multiwords that are not commonly used in general language; Verberne et al. (2010) estimate the overlap of "two-word" terms occurring in patent text and general text to be fewer than 2%. This pattern is also reported by D'hondt (2009) who states that these [multiword] terms are invented and defined ad hoc by the patent writers, and will generally not end up in any dictionary or lexicon. Understanding what these terms mean can be a formidable task: Turk (2006: 43) notes, "while determining what the claims mean may sound relatively simple, it is a task that federal district courts, the Court of Appeals for the Federal Circuit (CAFC), and even the United States Supreme Court have struggled with for years. Invention complexity, linguistic interpretation, and shifting court rules have kept inventors, the USPTO, and patent attorneys guessing as to what patent claims actually cover."

Importantly, the above issues also open up the possibility for patent writers to strategically omit relevant information to make claims appear more novel than they actually are, or include irrelevant information. The Government Accountability Office Report (GAO, 2016) estimated that 64% of examiners found excessive references made it difficult to complete prior art search in the allotted time, and 88% of examiners reported consistently encountering irrelevant references. Prior research has also documented evidence of strategic patent applications: Alcacer and Gittelman (2006) show *two-thirds* of prior art citations in an average patent are inserted by examiners to define appropriate scope, thus documenting that patent applicants often fail to include relevant information on prior art. Moreover, Lampe (2012) finds that applicants withhold between 21 and 33% of relevant citations, also suggesting applicants are strategic in who they cite.

The key challenge in the review of a focal application is then to discern among two plausible explanations for its claims. The first explanation is that there is no strategic altering—the patent claims are truly novel and representative of prior art, and the content of the application has all the relevant information for complete inputs in the search process. The second explanation is that the application lacks all relevant information, and suffers from input incompleteness.[8] The fact that both explanations are likely, but not certain, creates the potential for Type 1 and Type 2 errors.

3.2 | Technologies for search and prediction of relevant prior art

First deployed in the late 1990s (Dickinson, 2000), the current EAST (Examiners' Automated Search Tool) and WEST (Web-based Examiners Search Tool) systems use Boolean operators to conduct structured searches using keywords and strings with all available databases to yield potential prior art citations. GAO (2016) noted the current search tools were a significant constraint on productivity: More than half of the patent examiners thought these search tools were "less advanced and more reliant on keyword searching than other available tools" (p 24). The report stated the majority expressed a desire for increased automation

[8]We note input incompleteness may also arise because patent applicants are truly unaware of relevant prior art.

in search engines, both in search for concepts and synonyms related to examiner entered keywords, and in identification of prior art based on information already included in the patent application (claims and specification) without reliance on patent examiner entered keywords.[9]

4 | ML TECHNOLOGY AND HUMAN CAPITAL COMPLEMENTARITY IN PATENT EXAMINATION

4.1 | Development of ML technology for patent examination

To address these challenges, the USPTO has invested resources in developing a ML tool called *Sigma* (Krishna et al., 2016). While patent examiners are the first and critical stakeholders, as with the earlier vintages, USPTO intends to make ML tools public to benefit the population of potential inventors and personnel employed in the patent application process.

Over time, the USPTO team trained six versions of classification algorithms, prior to selecting the final algorithm incorporated in the *Sigma* tool (Krishna et al., 2016).[10] To use *Sigma*, examiners manipulate the search algorithm by selecting patent application components (e.g., abstract, title, description) they believe are most relevant, and enter the search terms into a "word cloud."[11] *Sigma* in turn adjusts the algorithm and retrieves a more refined set of search results in two conceptual steps. First, the algorithm starts with a *baseline set of search terms from the claim text being examined* and uses a classification-based ML approach to identify an expanded set of relevant search terms from training documents. These intermediate prior art documents are termed "pseudo relevance feedback" (PRF) documents. The classification algorithms find additional query terms that co-occur frequently with the initial query terms. Second, language-modeling techniques are used to select terms from the PRF documents and add such terms to the set of query terms. The expanded set of query terms are then employed to search for relevant prior art. In the baseline case, these ML algorithms make a prediction of relevant prior art *without user intervention*. Moreover, *Sigma* learns the examiner's habits to suggest searches based on their prior behavior.

4.2 | ML technology and human capital complementarity

We posit two human capital attributes—domain-specific expertise and vintage-specific skills—are complements to ML technology. Prior to ML technology to the cognitive tasks of identifying and rectifying inherent potential for input incompleteness rested squarely upon patent examiners. They then searched for prior art through manual reading and cataloging (prior to digitization), or using Boolean search (post digitization).

[9]An additional problem is that existing search technologies retrieve only a small fraction of relevant prior art. In part, this is because tools based on textual distance measure between the focal application and potential prior art often favor larger term frequencies and thus longer documents among the prior art corpus (Bashir and Rauber, 2010).

[10]See Appendix B Exhibit B4 for a visual of the interface. To train the algorithm, the USPTO created a test corpus of 100,500 patent documents with 60,300 granted patents and 40,200 pre-grant publications. Specialists conducted searches across the different technical domains and handpicked references made available for testing (Krishna et al., 2016).

[11]Word clouds visualize text data through increases in size of search terms deemed more important than others.

A central question of ML is whether domain-specific expertise will retain any value when machines can learn. Indeed, if domain expertise could be fully incorporated into the training dataset and algorithms used to generate predictions, and the challenges in the patent examination process were entirely due to constrained productivity of search engines, then perhaps machines could adjudicate patents (mostly) on their own.[12] However, the challenge posed by the dynamic updating of language in patent applications discussed above creates (strategic) input incompleteness when the search is based solely upon the information in the focal patent application. In theory, an ML algorithm may be able to overcome some of these challenges, as long as the strategic omissions/inclusions in the language of patent text are static. But, given that novelty and nonobviousness of true inventions also require new language to be constructed, and old language to become irrelevant, the future itself is unfamiliar terrain. Such dynamism inherent in the patent examination context makes it exceptionally difficult for ML to make reliable predictions about a future that is unfamiliar to its training dataset. Domain experts, however, may not be as susceptible to input incompleteness, because they will bring relevant tacit information and judgment based on prior knowledge into the search process. Thus, they will be better able to identify the difference between true shifts in the knowledge base and strategically written applications.

Second, the *potential* of ML technology's productivity improvement can be *realized* contingent on individuals interfacing with the technology having vintage-specific human capital. Here, we note our focus is on workers not as *developers*, but *users* of ML technologies. We posit workers with computer science and engineering (CS&E) knowledge bases will be more productive with ML technologies relative to those with non-CS&E knowledge bases as the user-interface of ML technologies will be more familiar to them. In their seminal article, Cohen and Levinthal (1990) note an individual's prior related knowledge is key to absorptive capacity. Prior computer programming knowledge provides CS&E workers a better skill-set for interfacing with ML tools, relative to non-CS&E workers who may lack these skills and expertise.

Put simply, our hypothesis is that ML will not address biases due to input incompleteness without complementary domain specific expertise, and user-interface complexities of ML require that humans who provide such expertise also have complementary vintage specific human capital.

5 | OBSERVATIONAL EVIDENCE

We use both observational and experimental evidence to provide support for our main propositions (see Table 1). In this section, we bolster two assumptions implicit in our claims. First, there is dynamic updating of language in the underlying knowledge base (potentially due to true novelty or strategic input incompleteness). Second, patent examiners with more domain expertise do indeed bring in more external information when devising search strategies for prior art.

[12]As noted above, ML algorithms do not yet have a comparative advantage over humans in cognitive tasks requiring tacit knowledge, flexibility, judgment and creativity (Agrawal et al., 2018; Autor, 2014), and this is also what distinguishes experts from novices (Simon, 1991).

TABLE 1 Summary of empirical analyses

Core proposition and summary	Observational or experimental test?	Results summarized in
1. Patent language dynamically changes over time. This finding is important because it bolsters the assumption that ML will have difficulty finding the most relevant prior art when the patent text itself includes words not found in the training data.	Observational test	Figure 1
2. More experienced examiners bring in additional knowledge to the adjudication process. This analysis bolsters a key assumption underlying the hypothesized complementarities between ML and domain-expertise. If examiners did not bring in external knowledge in the patent adjudication process, then there would be no basis to claim that domain expertise could address input incompleteness.	Observational test	Table 2
3. Relative to the Boolean vintage, machine learning search technology directs examiners to a narrower set of patents. This finding establishes the potential value of ML technology, because it could potentially find the most relevant prior art, thus saving examiners and the USPTO precious time and resources.	Experimental test	Figure 2, Table 5
4. Relative to the Boolean vintage, machine learning search technology directs examiners to patents that are much more textually similar to the focal patent application, but less similar to the silver bullet patent. This experimental evidence bolsters the point that ML can find the most textually similar prior art, but that it may not be the most relevant prior art.	Experimental test	Figure 3, Table 6
5. Domain expertise improves the likelihood of finding the silver bullet for both technologies, and it has a stronger effect in redirecting those using ML away from patents that are textually similar to the focal application. This finding documents that domain expertise shifts the narrow distribution closer to the most relevant prior art.	Experimental test	Figures 3, 4 and Tables 6, 7
6. Those with CS&E backgrounds perform better on ML technology, and this differential is driven by their higher ability to handle user interface complexities when implementing expert advice.	Experimental test	Figures 4, 5 and Table 8

5.1 | Dynamic updating of language describing knowledge base

While prior literature has alluded to dynamic updating of language, we examine whether this is truly the case by examining trends in a narrow knowledge domain. Specifically, we document temporal changes in the language structure in patents by collecting textual data related to all 949 patents in the subclass of the focal patent in our experimental setup (described in the next

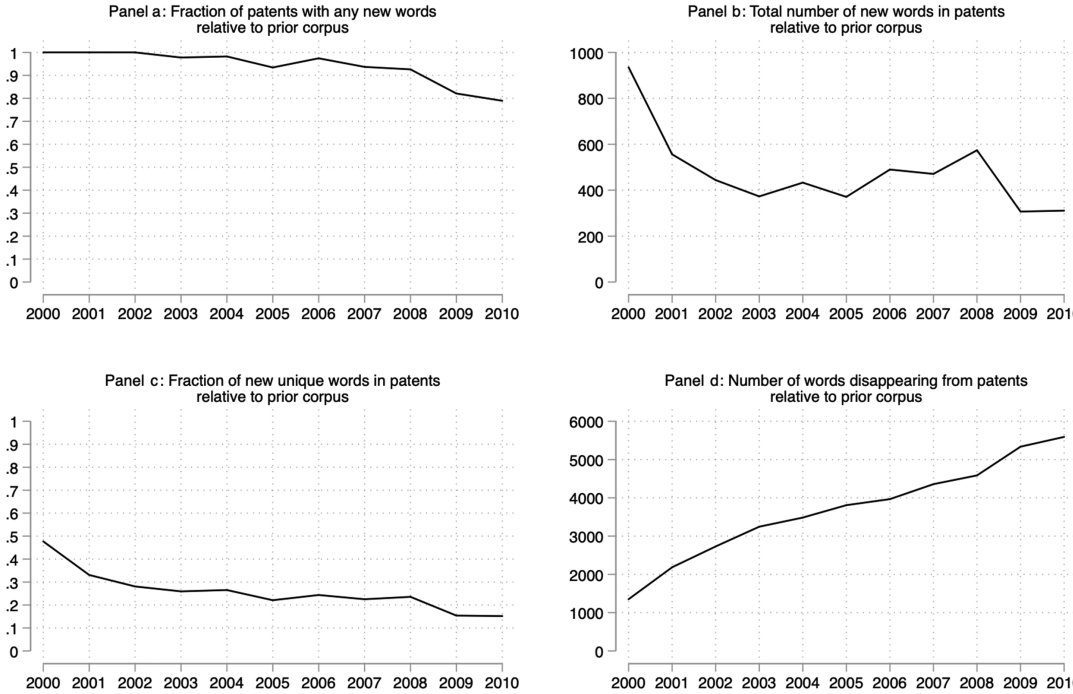

FIGURE 1 The changing nature of language in patent text

section) from 1995 to 2010.[13] Within each patent application, we extracted all unique words in the title, abstract, and first claim,[14] and then aggregated to the year level, to obtain a database of all the unique words in each patent in a given year. We further aggregated the corpus of unique words in this patent class for an initial period of five years, and then analyzed how language in patents changed over time relative to the prior corpus of unique words (which accumulates year over year).

Panel a of Figure 1 documents the proportion of patents with *any* new words relative to the prior corpus. Even after 15 years of collecting unique words, approximately 80% of patents still have at least one new unique word. Panel b of Figure 1 tracks the number of unique words in each year relative to the prior corpus, while Panel c tracks the ratio of new unique words added to the patent corpus each year, measured as the number of *new* unique words in patent text filed each year divided by the total number of unique words in annual patents. As would be expected, both plots display a downward trend, so that there fewer new words in patents over time, but both are persistently positive. Lastly, instead of examining *new words*, Panel d examines the number of unique words that were previously in the corpus of words but did not appear in any patents in a given year. It shows that more and more words *disappear* over time relative to the prior corpus. Altogether, Figure 1 documents two substantive challenges to USPTO examiners and the technology they utilize: new language is added to the text of patent claims and old language disappears from use.

[13]The specific subclass is subclass 46 of the United States Classification (USPC) main class 725 (Interactive Video Distribution System). https://www.uspto.gov/web/patents/classification/uspc725/defs725.htm. The patent data were drawn from the official USPTO website's bulk data product, PatentsView (http://www.patentsview.org/download/).
[14]The first claim is usually the most important. The data are accessed from https://www.uspto.gov/learning-and-resources/electronic-data-products/patent-claims-research-dataset

This analysis bolsters a key assumption related to the viability of ML technologies: If the language of patent applications is constantly changing, whether because of novel and nonobvious knowledge, or because of deliberate attempts to omits or add unnecessary language, ML tools that operate by reading the patent application may struggle to find the most relevant prior art.

5.2 | Relevance of domain expertise in patent examination process

To examine whether experienced patent examiners are more likely than novice examiners to bring information *not* present in the focal application to bear in the adjudication process, we use data drawn from the official USPTO website and their bulk data product, PatentsView. We first randomly sampled 15,000 patent applications using the Patent Examination Data System (see https://ped.uspto.gov/peds/). We then extracted all available patent examiners' SRNT files, which contain the *actual keyword searches* used by examiners to search for prior art for a given patent application, giving us 11,050 patent examiners' SRNT files from the years 2001 to 2012.[15] We also collected data on patent abstracts, titles and United States Classification from PatentsView as well as information on patent first claims from the USPTO patent claims dataset.

Based on these data, we constructed two variables to analyze the relationship between *examiners' contributions* of keywords to search for prior art and their *experience*. Examiners' contributions, our dependent variable, was measured as the percentage of new, unique keywords added by examiners in their search for prior art (as coded in the SRNT file) compared with the unique words in the patent's abstract, title and first claim. That is, the dependent variable measures the number of unique keywords used in the search process that were *not present* in the patent text being examined, scaled by the number of unique words in the patent text. Examiner experience, our independent variable, was measured as the number of years following the first patent examination.

Table 2 reports OLS results examining the relationship between experience and examiner contributions, without (columns 1–2) and with (columns 3–4) examiner fixed effects. In the even columns, we drop the first 2 years of patent examination because in this period novice examiners are almost always paired with more senior examiners to assist in identifying prior art. Standard errors are clustered at the examiner level. We find that, in each case, an additional year of examiner experience is associated with an increase of 0.007 to 0.015 examiner contributions, such that an examiner with 10 more years of experience has 10–20% more examiner contributions on average.

This analysis bolsters the central mechanism that domain expertise is relevant to address input incompleteness, because it shows examiners bring in more external words (i.e., words not found in the application text) into the search process as they gain experience.

6 | EXPERIMENTAL ANALYSIS

Our experimental design enables us to examine whether and how domain expertise and CS&E knowledge complements ML technologies to mitigate biased predictions in the face of input incompleteness, when compared to the existing Boolean search technology.

[15]SRNT files represent "search notes," that is, notes on searches performed by the examiner during the examination of an application. (https://www.uspto.gov/sites/default/files/patents/process/status/private_pair/PAIR_Glossary.pdf). The SRNT files were converted to processable text using pytesseract (https://pypi.org/project/pytesseract/).

TABLE 2 Examiner-added search terms and examiner experience

Dependent variable	(1) Examiner's additional search contribution relative to patent application	(2)	(3)	(4)
Years of experience	0.011	0.015	0.007	0.011
	(.036)	(.007)	(.075)	(.016)
Constant	0.696	0.664	0.714	0.688
	(.000)	(.000)	(.000)	(.000)
Observations	11,050	9,871	11,050	9,871
R-squared	0.011	0.019	0.292	0.299
Sample	All	Experience>2 years	All	Experience>2 years
Examiner FE	No	No	Yes	Yes

Note: p-Values in parentheses, with standard errors clustered at the examiner level. The dependent variable is the ratio of the unique search words the examiner searched for that did not appear in the patent title, abstract, or first claim, divided by the number of unique words in the patent title, abstract, or first claim. The mean of the dependent variable is 0.75 in both the full sample and in the sample with experience of more than 2 years.

6.1 | Description of experiment

The ideal design to examine productivity differentials of a process technology based on interactions with heterogeneous human capital is one in where (a) subjects do not self-select into the experiment; (b) technology vintage type is randomly assigned across individuals with and without CS&E knowledge, and domain expertise is randomly manipulated; (c) there are ways to assess potential contamination between treatment and control group; and (d) it is possible to not only causally identify the relationships of interest, but also shed light on the underlying mechanisms. We briefly describe the experiment, and then highlight how our research design allows us to address each of these concerns (see Appendix B, Experimental Design Section).

The participants in this experiment were 221 graduate students enrolled in three sections of a course at a top tier business school. They were asked to "examine" a patent with five claims over a period of 5 days. The choice of MBA students as experimental subjects allowed us to control for unobservable differences across subjects in human capital (e.g., domain expertise), since all subjects had a similar lack of experience with the patent examination process and the search technologies used within them. Accordingly, they better simulate novice examiners (with differences in prior CS&E knowledge bases) being assigned to different technologies.[16]

The research team simulated conditions encountered by patent examiners in designing the patent examination exercise with invaluable help from USPTO officials. First, we determined the time period to be allotted for examination of the patent claims, and also ensured task manageability given expectations of time investments by the subjects in the experiment.[17] Second,

[16]In addition, as noted above, the USPTO intends both technologies to be available for general public use, and the heterogeneity in prior backgrounds of MBA students may be more representative of the skilled labor force at large and thus informative of productivity differences for larger stakeholder groups.

[17]Given that the examiner spends on average 18 hr examining the claims in a patent (Lemley & Sampat, 2012), and that an average patent has around 14.9 claims (Allison & Lemley, 2000), the examiner spends around 0.4 hr to examine a single claim. In consultation with USPTO officials, we determined 0.6 hr per claim to be reasonable; the additional 33% of time allotment per claim provided a cushion given the novelty of the task to the participants.

USPTO officials helped identify a past patent application of average complexity for use in the experiment. The specific patent application had been examined in 2010, and was rejected based on a *single* prior art citation—the "silver bullet"—that invalidated its novelty claims (Appendix B, Exhibits B5 & B6). To reflect the potential strategic gaming of patent applications, the silver bullet patent is not too similar to the focal application. Given that the applicant had abandoned the process after the claims were rejected, no records of the patent examination results were available to our experimental participants (e.g., as could be identified through a Google search). To pare down the time investments to manageable levels, the USPTO officials identified 5 claims from the 39 claims made in the original application which were deemed most relevant for the rejection.

Third, to enable random provision of "domain expertise," an experienced USPTO examiner in the art-unit that originally examined the patent in 2010 created the content of the emails that were shared as "expert advice." This expert advice (which codifies domain expertise, see Appendix B, Exhibits B8 & B9) included keywords to be used to search for relevant prior art, including several words from the text of the patent being examined (e.g., "data," "program," "watch," "watching," etc.) and several keywords *not present in the text of the patent being examined* (e.g., "viewer," "monitor," "recording," "tuning," etc.). The expert advice was additionally tailored to the process technology, ensuring that following the steps outlined in the advice resulted in the best performing search strategy for each technology. Thus, the expert advice was both task and technology specific, and recommended explicit search strategies on manipulating a word cloud to insert examiner added keywords (for the ML group) and composing the search string with the examiner added keywords (for the Boolean search group).

Fourth, postsearch adjudication was simplified such that identification of the "silver bullet"—one prior patent which would invalidate all claims in the focal patent application—would result in the correct decision, that is, that the application be rejected.[18] Finally, the university information technology department worked with USPTO officials to ensure smooth functioning of both technologies and capture the work-logs of participants as they interfaced with the technology.

The actual experiment proceeded over 2 weeks. In Week 1, participants received training on the patent examination task by officials from USPTO, and were provided information similar to new patent examiner hires. Here, they were introduced to *both* search technologies, but at this point, participants were unaware of which technology would be assigned to them. At the end of Week 1, participants were randomly assigned to a technology—Boolean or ML—using a random number generator. Each participant received additional training materials to ensure familiarity with the technology and task, and had 5 days to complete the exercise. The participants then initiated the exercise, searching for prior patent references. Half of the participants on each technology were randomly chosen to receive the email containing expert advice from the USPTO patent examiner in the midst of this exercise. Correctly inputting these strategies resulted in a list of the prior art potentially relevant to the specific patent application; however, it was not sufficient for task completion. The subject had to read through the output and

[18]Our field interviews reveal that patent examination often entails the identification of a "silver-bullet" (this is a term used by patent examiners) to invalidate claims being examined. Expert examiners often identified the silver-bullet through a process described to us as "rummaging," that is, using different relevant keywords to identify relevant prior art. Once the examiner is able to identify a silver-bullet, they often search for other prior art references that might invalidate the claims. This is done by searching for prior art in the relevant technology sub-class. The examiner is assured of the existence of a silver-bullet if all relevant search pathways led to a single patent (i.e., the silver-bullet) that could invalidate the claims.

TABLE 3 Summary statistics

	Full sample		Sample that received expert advice		Sample that did not receive expert advice	
Variable	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
1(Silver Bullet Cited)	0.06	0.24	0.12	0.32	0.00	0.00
Times Silver Bullet Cited	0.10	0.51	0.21	0.71	0.00	0.00
1(Machine Learning)	0.50	0.50	0.50	0.50	0.50	0.50
1(Expert Advice)	0.50	0.50	1.00	0.00	0.00	0.00
1(CS&E) (computer scientists & engineers)	0.22	0.42	0.23	0.42	0.22	0.41
Time spent on tool (min)	22.62	19.89	21.71	19.08	23.51	20.71
1(Fluent in English)	0.57	0.50	0.59	0.49	0.55	0.50
1(Section 1)	0.41	0.49	0.41	0.49	0.41	0.49
1(Section 2)	0.31	0.46	0.32	0.47	0.31	0.46
1(Section 3)	0.28	0.45	0.27	0.45	0.28	0.45
Observations	221		110		111	

determine whether the prior art references invalidated the claim. At the end of the exercise, we assessed which patents were cited, the similarity in those patents to the "silver bullet" and focal patent application, the finding of the silver bullet, and examiner speed.

As it relates to the four criteria for an "ideal experiment," first, participants did *not* self-select into this experiment. All participants were students enrolled in a semester-long, second-year elective MBA course, and they were unaware at time of enrollment of their potential participation in the experiment. While participation in the experiment was voluntary, two of the three sections of the course had a 100% participation rate for students physically on campus that week,[19] and the third had a participation rate of 81%. In addition, participants signed consent forms prior to knowing the technology they were assigned, alleviating some concern that participation might be endogenous to being assigned to a technology. Second, to prevent across group contamination, participants were asked to sign an academic declaration of honesty, stating they would not collaborate with anyone during the exercise. Moreover, access to the electronic activity logs enables us to test how many of the individuals in the group that *did not* receive expert advice ended up using the advice as suggested by the e-mail: we find that this occurred in only three cases, and the results are robust to these cases being excluded from the analysis. Finally, the activity logs for each participant also help shed light on the underlying mechanisms driving the causal relations of interest, including whether or not participants integrated the expert advice into their search activities.

Table 3 presents descriptive statistics for the sample. Overall, the average participant spent 22.6 min on the tool.[20] It is important to note that the minutes spent on the tool only include when the examiner was logged in. Examiners were logged out automatically after 90 s of inactivity on the tool, and had to re-enter their search queries when they logged back in.

[19]For these two sections, when counting all students including those not physically on campus, the participation rate was 97% and 94%, respectively. The students who were not on campus were traveling out of town for job interviews.
[20]Twenty-three participants spent no time working on the tool, and they seem to be randomly distributed across groups. Also, neither technology, nor CS&E background, nor expert advice is predictive of the time spent on the time tool.

6.2 | Variable definitions

6.2.1 | Dependent variables

Success in the task is based on whether the patent application being reviewed was rejected for all five claims due to participants identifying *the silver bullet*. To measure an examiner's success, we count the number of times they cited *the silver bullet* in their evaluation of each of the five claims. In robustness checks, we use as a dependent variable an indicator for citing the silver bullet patent at all. To measure an examiner's *productivity*, we compute the ratio of silver bullet citations to time spent on the platform. We also examine dependent variables that reflect Levenshtein similarity scores (Thoma et al., 2010) between the patents cited by the examiners and both a) the focal patent, and b) the silver bullet patent.

6.2.2 | Independent variables

We manipulated two variables in this experiment: First, ML_i is a dummy variable set to 1 if examiner i was assigned to use the ML based process technology, Sigma (and 0 otherwise). Second, $Expert\ Advice_i$, is a dummy variable set to 1 if examiner i received the expert advice email (and 0 otherwise). The prior knowledge base of examiners is captured by the indicator variable $CS\&E_i$, set to 1 if examiner i has a degree in computer science and engineering (and 0 otherwise). The variable $Expert\ Advice_i$ encapsulates the construct of domain-specific expertise; on the other hand, the variable $CS\&E_i$ reflects vintage-specific human capital for ML technology.

6.2.3 | Controls

Although randomization ensures unbiased estimates, we include several pretreatment control variables to reduce residual variation and increase the precision of our estimates. These include indicators for section, gender, whether the individual has a partner, and U.S. citizen.

6.3 | Balance tests and empirical approach

Each examiner was randomly assigned to one of four treatment groups (Boolean vs. ML; Received Expert Advice vs. not), stratified by examiner background (CS&E skills vs. not) for each class section.[21] Table 4 presents means and standard errors of various pretreatment variables by treatment group. The stratification and matching seem to have worked as designed because the distribution of CS&E and section are evenly distributed between the four treatment groups. Other pretreatment variables are also balanced across the groups (see Table 4).

[21]The sampling method comprised three steps. First, we sorted student IDs by section and CS&E background. Second, we matched subjects in the four treatment groups to the six strata (3 sections*CS&E background dummy). This ensured each of the six strata have an approximately equal number of subjects assigned to each treatment group. In the final step, within each stratum, we shuffled the treatment assignments using a random number generator. We verified our randomized treatment assignment generated roughly equal numbers of each treatment within each stratum.

TABLE 4 Balance checks

	Boolean	Boolean + expert advice	Machine learning	Machine learning + expert advice	*p*-Value of joint test of equality
1(CS&E)	0.196	0.218	0.236	0.236	0.952
	(.05)	(.06)	(.06)	(.06)	
1(Male)	0.625	0.473	0.582	0.509	0.368
	(.07)	(.07)	(.07)	(.07)	
1(Has partner)	0.268	0.436	0.309	0.309	0.264
	(.06)	(.07)	(.06)	(.06)	
1(U.S. citizen)	0.571	0.582	0.636	0.6	0.907
	(.07)	(.07)	(.07)	(.07)	
1(Section 1)	0.411	0.418	0.418	0.4	0.997
	(.07)	(.07)	(.07)	(.07)	
1(Section 2)	0.304	0.309	0.309	0.327	0.994
	(.06)	(.06)	(.06)	(.06)	
1(Section 3)	0.286	0.273	0.273	0.273	0.998
	(.06)	(.06)	(.06)	(.06)	
Observations	56	55	55	55	

Note: This table reports the mean and standard deviations of the variables in the row across our four experimental subsamples. The fifth column is the *p*-value of the joint test of equality of the means.

To examine the causal effects of ML and the random provision of expert advice, we estimate the following set of equations using OLS[22]:

$$Y_i = \gamma_0 + \gamma_1 ML_i + B\boldsymbol{X} + \varepsilon_i \quad (1)$$

$$Y_i = \alpha_0 + \alpha_1 Expert\ Advice_i + B\boldsymbol{X} + \varepsilon_i \quad (2)$$

$$Y_i = \beta_0 + \beta_1 ML_i + \beta_2 Expert\ Advice_i + \beta_3 Expert\ Advice_i * ML_i + B\boldsymbol{X} + \varepsilon_i \quad (3)$$

where $\boldsymbol{X}$ includes the controls noted above and Y_i is either the number of times the silver bullet is referenced, the examiner's productivity (silver bullet citations per minute), or the similarity between the patents cited and the silver bullet or focal patent applications. Equation (1) estimates the causal effect of ML relative to the Boolean technology, while Equation (2) examines the causal effect of the provision of expert advice, and Equation (3) examines the interaction between expert advice and ML. In subsequent models we add a third interaction of CS&E.

[22]We use OLS rather than count models (e.g., Poisson) because many observations are perfectly predicted, as is apparent in Figures 3 and 4. The results are robust to LPM models, both in terms of magnitude and statistical significance.

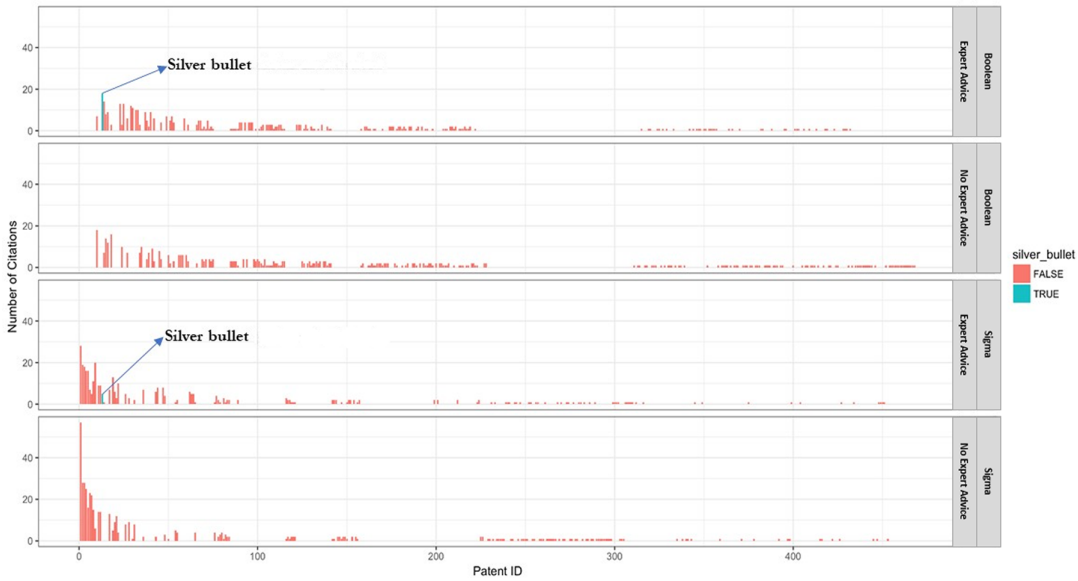

FIGURE 2 Distribution of patents cited by technology type and expert advice [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 5 Number of cited patents by technology and expert advice

	Total unique patents cited	75% citations from how many patents?
Boolean	204	92
Boolean + expert advice	158	64
Machine learning	122	27
Machine learning + expert advice	107	33
Average	148	54

7 | RESULTS

We begin our analysis by examining which patents were cited by the subjects based on randomly assigned technology and expert advice. Figure 2 depicts the full distribution of patents cited by these groups, where each point on the *x*-axis is a unique patent. The figure clearly shows examiners assigned to the ML technology cite a much narrower (more precise) distribution of patents than the Boolean technology. Table 5 provides the total number of prior art patents cited, and validates the distinction between the Boolean and ML groups evident from Figure 2: in the two Boolean treatment groups, there were 204 (without expert advice) and 158 (with expert advice) unique patents cited, versus 122 and 107 in the two ML groups respectively. Moreover, of all the citations made by the Boolean group, 75% came from 92 unique patents in the sample without expert advice, and 64 unique patents in the sample with expert advice. In contrast, 75% of the patents cited from the ML groups came from just 27 (no expert advice) and 33 unique patents (with expert advice). We performed two-sample Kolmogorov–Smirnov tests and reject the null that the distributions are not different from each other ($p < .05$). Similarly, chi-squared tests to see if the same patents were cited in each treatment

group also reject the null ($p < .05$). These patterns validate that ML based technologies lead to a narrower distribution.

While this analysis sheds light on the potential value of ML in the patent examination context, it does not give a sense of how those cited patents might actually be different in ways that help us understand the role played by ML and expert advice. We supplement the prior analysis by comparing the text of each patent cited to both the focal patent application and the silver bullet patent. In particular, we calculate the Levenshtein distance between each patent cited by each examiner and both the silver bullet patent and the focal patent application. This similarity measure is bounded between 0 (least similar) and 1 (most similar).[23] Table 6 reports results from estimating Equations (1)–(3) with similarity to the silver bullet in Panel a, and similarity to the focal patent application in Panel b. Figure 3 summarizes these results by plotting average similarity for each of the four groups to the focal patent application and silver bullet.

Figure 3 and Table 6 show that without expert advice, the ML technology pushes examiners toward patents that are more similar to the focal patent application (Panel b columns 7 and 8). In this sense, ML appears to be doing what it is meant to do: utilizing natural language processing, the ML technology "reads" the text of the focal patent better than the Boolean technology, generating predictions that more accurately reflect prior art that appears in the *content* of the focal patent application. However, the text in the silver bullet patent is actually not that similar to the text of the focal patent application. For example, without expert advice, the mean similarity between the cited patents on the ML technology and the focal patent application is approximately 0.25, when in fact the silver bullet has a similarity of only 0.09. As a result, those on the ML technology cite patents that look more like the focal patent, but less like the silver bullet patent relative to those using the Boolean technology (column 2 and 8 of Table 6).

Due to the dissimilarity between the focal patent application and the silver bullet, additional keywords in the search help to identify the silver bullet. Such knowledge is often tacit and is only possessed by human experts, and thus it is perhaps not surprising that expert advice helps examiners using both technologies by redirecting their search efforts toward the silver bullet. This can be seen visually via the *upward* shift in Figure 3, which is substantiated in columns 5 and 6 of Table 6: Expert advice increases the similarity to the silver bullet for both Boolean and ML technologies, and there is no statistically distinguishable difference in the extent of the increase. In contrast, columns 11 and 12 of Table 6 show that while expert advice pulls the Boolean technology to less similar patents relative to the focal application, it pulls the ML examiners even further (i.e., the leftward shift associated with Expert Advice in Figure 3 is greater for ML than Boolean).

In summary, this similarity analysis clearly shows that ML is finding patents very similar to the patent application, but that expert advice is even more necessary to redirect examiners on ML toward the patents that look more like silver bullet than on the Boolean technology. However, finding patents that look like the silver bullet—at least in this case—is no substitute for finding the actual silver bullet. We now turn to that analysis.

Table 7 reports our main results with regards to the effects of ML and expert advice on the number of times the silver bullet is cited, as well as the number of silver bullet citations per minute (productivity). As the similarity analysis suggested, columns 1 and 2 of Panel a show ML is actually associated with 0.12 fewer citations of the silver bullet. ML is not necessarily less efficient in terms of time (columns 7 and 8 of Panel b), indicating the gains in finding the silver bullet on the Boolean technology (from Panel a) came at the cost of more time. Columns 3 and 4 of Panel a show that, as expected, expert advice causes the silver bullet to be cited 0.21 more times,

[23]Since examiners cite multiple patents, we use the within-examiner average across all cited patents.

TABLE 6 Similarity between the cited patents, the silver bullet, and the focal patent

Model: OLS	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Dependent variable:	**Panel a: Similarity to silver bullet**						**Panel b: Similarity to patent application**					
1(Expert Advice)			0.025	0.024	0.034	0.033			−0.039	−0.037	−0.017	−0.014
			(.000)	(.001)	(.001)	(.004)			(.002)	(.005)	(.001)	(.015)
1(ML)	−0.014	−0.014			−0.003	−z0.004	0.127	0.125			0.142	0.142
	(.060)	(.075)			(.251)	(.215)	(.000)	(.000)			(.000)	(.000)
1(Expert Advice)*1(ML)					−0.019	−0.018					−0.034	−0.036
					(.185)	(.228)					(.026)	(.022)
1(Male)		−0.009		−0.005		−0.004		0.013		0.006		0.009
		(.267)		(.547)		(.578)		(.127)		(.629)		(.266)
1(Has Partner)		0.006		0.004		0.001		−0.013		−0.024		−0.010
		(.529)		(.651)		(.876)		(.155)		(.086)		(.235)
1(U.S. Citizen)		0.010		0.007		0.008		−0.003		0.008		−0.002
		(.238)		(.379)		(.310)		(.729)		(.539)		(.841)
1(Section 2)		−0.005		−0.005		−0.005		−0.004		−0.004		−0.003
		(.673)		(.620)		(.622)		(.686)		(.809)		(.755)
1(Section 3)		−0.002		0.000		−0.000		−0.007		−0.015		−0.007
		(.869)		(.996)		(.984)		(.514)		(.354)		(.449)
Constant	0.112	0.111	0.092	0.091	0.093	0.093	0.100	0.104	0.184	0.189	0.109	0.112
	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)
Observations	163	163	163	163	163	163	163	163	163	163	163	163
R-squared	0.022	0.040	0.073	0.083	0.101	0.110	0.601	0.612	0.058	0.082	0.655	0.662
Sample	All	All	All	All	All	All	All	All	All	All	All	All

Note: *p*-Value in parentheses, with robust standard errors. ML stands for the machine learning technology.

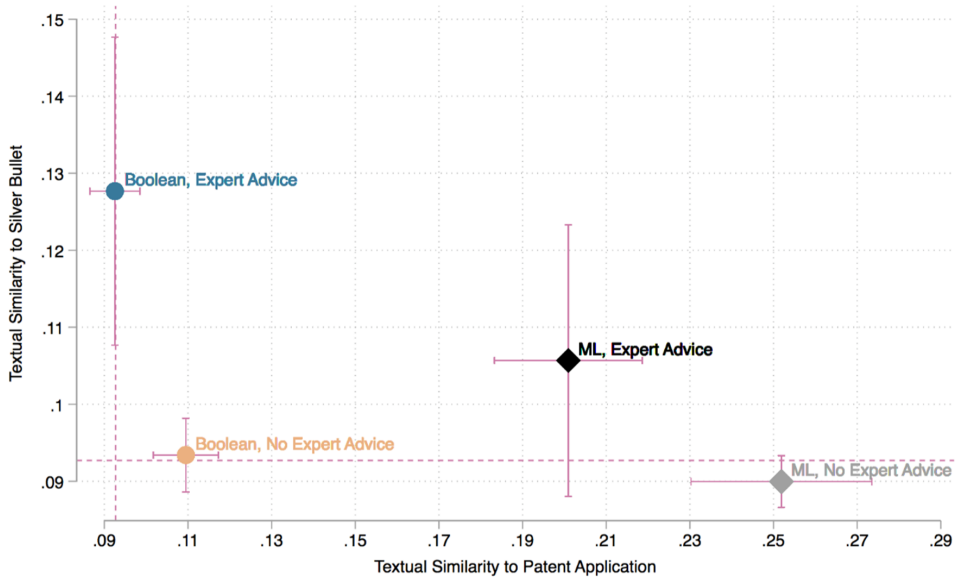

FIGURE 3 Mean similarity to patent application and silver bullet by technology, expert advice [Color figure can be viewed at wileyonlinelibrary.com]

increasing productivity by twice the sample mean. Considering the interaction of expert advice and ML, the point estimates suggest that expert advice is strongly associated with finding the silver bullet on the Boolean technology (0.327 in column 5), but is relatively less effective with ML (−0.236 in column 5), though the negative interaction is not robust to including controls.

7.1 | Vintage-specific capital, absorptive capacity, and robustness checks

Thus far, we have documented the need for domain expertise to redirect ML algorithms to find the right prior art. But a second barrier to productivity, as with any new technology, is the know-how to interact with the technology. Here, we consider the role of prior CS&E background for effective interfacing with an early generation of the ML tool. Table 8 and Figure 4 report our main results. The only combinations of knowledge background and technologies able to identify the silver bullet were non-CS&E examiners working with the Boolean technology, and CS&E examiners working with ML technology.

To elucidate the role of absorptive capacity and CS&E background within the sample of those who received expert advice, we review the activity logs of examiners. In the ML subsample, there were three particular expert tips that needed to be followed to have a high probability of finding the silver bullet. The first two tips were relatively straightforward to integrate into the user interface: they included setting the weights and checking the "Claims" box. We call these the simple expert tips. The third tip, which is somewhat more complicated, involved manipulating the word cloud within the user interface—we call this the complex tip. Panel b of Figure 5 presents an unconditional binned scatter plot (dividing the *x*-axis into bins, showing the mean within each bin) of the number of times the search on the ML tool included the simple tips per minute spent on the tool. Visually, the slopes are quite similar between CS&E and non-CS&E, and column 4 of Table A3 confirms there is no statistical discrepancy between the

TABLE 7 The random provision of machine learning and expert advice

Model: OLS	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Dependent variable:	Panel a: Number of times silver bullet is cited						Panel b: Productivity (silver bullet citations per minute)					
1(Expert Advice)			0.209	0.206	0.327	0.325			0.007	0.008	0.011	0.012
			(.002)	(.003)	(.003)	(.006)			(.004)	(.008)	(.007)	(.016)
1(ML)	−0.117	−0.115			0.000	−0.001	−0.004	−0.004			−0.000	0.000
	(.087)	(.097)			(.000)	(.893)	(.100)	(.117)			(.000)	(.834)
1(Expert Advice)*1(ML)					−0.236	−0.233					−0.008	−0.008
					(.077)	(.104)					(.097)	(.121)
1(Male)		−0.043		−0.014		−0.008		0.000		0.001		0.001
		(.572)		(.852)		(.916)		(.982)		(.755)		(.670)
1(Has Partner)		0.044		0.027		0.008		0.000		−0.001		−0.001
		(.590)		(.735)		(.921)		(.980)		(.832)		(.658)
1(U.S. Citizen)		0.017		0.010		0.013		−0.001		−0.001		−0.001
		(.822)		(.893)		(.854)		(.811)		(.648)		(.690)
1(Section 2)		−0.049		−0.050		−0.049		−0.002		−0.002		−0.002
		(.529)		(.510)		(.517)		(.467)		(.449)		(.454)
1(Section 3)		−0.010		−0.005		−0.005		−0.002		−0.001		−0.002
		(.909)		(.955)		(.951)		(.575)		(.645)		(.612)
Constant	0.162	0.178	−0.000	0.012	−0.000	0.012	0.006	0.007	0.000	0.002	0.000	0.001
	(.005)	(.090)	(1.000)	(.852)	(1.000)	(.847)	(.009)	(.097)	(.000)	(.503)	(.000)	(.524)
Observations	221	221	221	221	221	221	198	198	198	198	198	198
R-squared	0.013	0.018	0.043	0.045	0.070	0.072	0.014	0.017	0.042	0.047	0.068	0.074
Sample	All	All	All	All	All	All	All	All	All	All	All	All

Note: p-Values in parentheses, calculated with robust standard errors. ML stands for the machine learning technology.

TABLE 8 CS&E complementarities with machine learning and expert advice

Model: OLS	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent variable	Panel a: Number of times silver bullet is cited				Panel b: Productivity (silver bullet cites/min)			
1(Expert Advice)			0.419	0.433			0.014	0.016
			(.003)	(.006)			(.007)	(.014)
1(ML)	−0.205	−0.203	0.000	0.000	−0.007	−0.008	−0.000	−0.000
	(.004)	(.007)	(.000)	(.970)	(.009)	(.016)	(.000)	(.939)
1(CS&E)	−0.205	−0.202	0.000	0.009	−0.007	−0.008	−0.000	−0.000
	(.004)	(.015)	(.000)	(.583)	(.009)	(.031)	(.000)	(.992)
1(ML)*1(CS&E)	0.397	0.392	0.000	−0.002	0.014	0.015	0.000	0.000
	(.021)	(.019)	(.000)	(.950)	(.017)	(.017)	(.115)	(.731)
1(ML)*1(Expert Advice)			−0.419	−0.430			−0.014	−0.016
			(.003)	(.005)			(.007)	(.012)
1(CS&E)*1(Expert Advice)			−0.419	−0.454			−0.014	−0.017
			(.003)	(.009)			(.007)	(.019)
1(ML)*1(CS&E)*1 (Expert Advice)			0.803	0.834			0.028	0.030
			(.017)	(.018)			(.015)	(.016)
Constant	0.040	0.043	0.125	0.129	0.007	0.010	0.000	0.002
	(.040)	(.043)	(.125)	(.129)	(.009)	(.069)	(.000)	(.376)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	221	221	221	221	198	198	198	198
R-squared	0.040	0.043	0.125	0.129	0.040	0.046	0.120	0.134
Sample	All	All	All	All	All	All	All	All

Note: p-Values in parentheses, estimated with robust standard errors. ML stands for machine learning, CS&E stands for computer science and engineering. Controls include gender, whether the individual has a partner, is a U.S. citizen, and indicators for the course section.

slopes. However, Panel c of Figure 5 shows clearly that non-CS&E examiners have much more difficulty incorporating the complex tip of manipulating the word cloud relative to CS&E examiners. Column 6 of Table A3 confirms the statistical differences in the slopes between the CS&E and non-CS&E examiners.

This analysis suggests one possible explanation for why non-CS&E do worse on ML, even with expert advice: They do not properly implement the expert advice into the ML interface, which then reduces the likelihood of their search surfacing the silver bullet. Indeed, CS&E examiners who found the silver bullet using the ML technology averaged 11.5 "expert" searches. In contrast, the non-CS&E examiners completed on average only 0.62 "expert" searches, with a maximum of seven.[24]

[24]Table A4, columns 4 and 6 confirm the heterogeneity between CS&E and expert searches matters for finding the silver bullet: column 6 shows that only CS&E examiners are able to translate expert searches into finding the silver bullet. Table A5 and A6 confirm the results are robust to examining simply finding the prior art and to productivity.

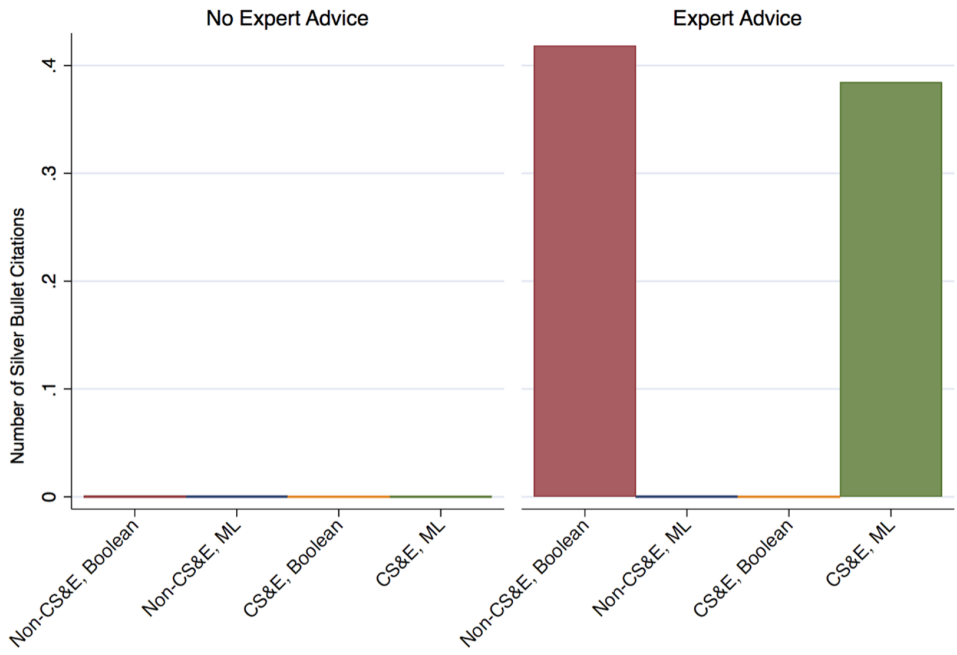

FIGURE 4 Silver bullet citations, CS&E background, and the receipt of expert advice [Color figure can be viewed at wileyonlinelibrary.com]

We next perform a similar exercise on the Boolean search tool, where the expert advice amounts to a simple copy and paste of the exact expression of the e-mail. Panel a of Figure 5 shows a binned scatter plot of the number of these "expert searches" performed based on the total number of minutes spent on the tool, broken out into CS&E and non-CSE&E categories.[25] The lines are essentially parallel, indicating that per-minute spent on the tool, CS&E and non-CS&E were equally effective in translating the expert advice into searches.[26] Despite being able to integrate the expert advice on the Boolean technology at the same rates, we see in column 2 of Table A4 that non-CS&E examiners are much more effective in turning those searches into finding the silver bullet.[27] We find this result somewhat puzzling, given that we would expect CS&E examiners to have equally good or better Boolean search skills than non-CS&E examiners.[28]

Whatever the explanation may be for why CS&E are worse on Boolean when they receive expert advice, one potential concern is that these results are driven by some spurious outcomes. We can provide some evidence this is not the case by showing the same patterns related to technology and CS&E are observed in both the sample of expert advice *and* the sample that did not

[25]Note that the respondents were automatically logged out after 90 s and when they logged back in they had to repeat their search. Minutes in the tool is the total amount of minutes the respondent spent logged in.

[26]Column 2 of Table A3 tests this comparison, confirming that there are no statistical differences in the slopes.

[27]Table A5 confirms the results hold when we examine productivity differences rather than just finding the silver bullet.

[28]We posit two potential theories for why we observe this distinction, though we are unable to further test them given data limitations: First, it may be the case that CS&E examiners, who were aware of the alternative ML technology, were discouraged that they were randomly assigned the Boolean search tool, and thus put forth less effort. That is, in line with Akerlof's theory of gift exchange (Akerlof, 1982), examiners who received the more desirable ML technology may have responded with additional effort relative to those who received the less desirable Boolean search technology. A second possibility is that CS&E examiners are less effective at parsing through large sums of reported text.

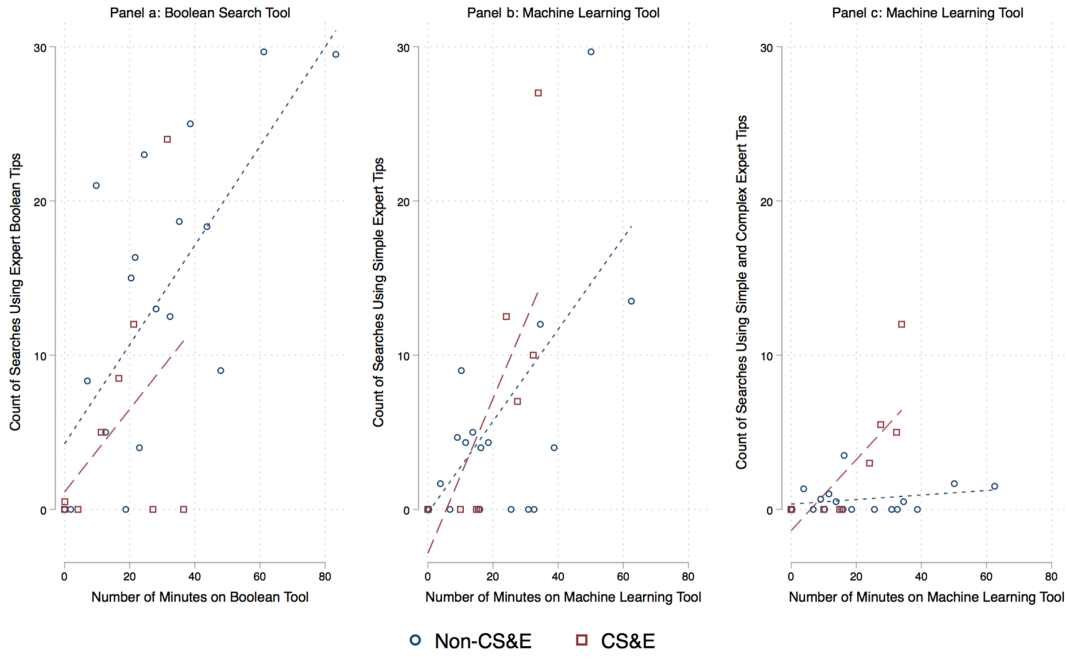

FIGURE 5 Incorporating expert search tips per minute, by CS&E background [Color figure can be viewed at wileyonlinelibrary.com]

receive expert advice. Table A7 uses as a dependent variable the similarity measure between the cited patents and the silver bullet,[29] and examines heterogeneous effects by technology and CS&E background, breaking out the effects separately by the receipt of expert advice. Although the magnitudes of the effects are different, the directionality and statistical patterns remain the same: Relative to a non-CS&E examiner on Boolean, CS&E examiners on Boolean cite patents that are less similar to the silver bullet, both in the sample with expert advice and in the sample without it. Similarly, non-CS&E examiners on ML cite less similar patents to the silver bullet than non-CS&E examiners on Boolean, in both the sample with and without expert advice, while CS&E on ML cite relatively more similar patents to the silver bullet (again, in both subsamples).

Given our experimental design randomizes technology vintage and expert knowledge, the results are likely robust to many alternative explanations and model specifications. Nonetheless, one potential concern is with the count measure as the dependent variable. We re-ran all our results with an indicator for whether an individual found the silver bullet citation at all, and the results, presented in the Appendix A, are qualitatively and quantitatively very similar (see Tables A1, A2, A5, and A6). Another potential concern is that voluntary participation in the study created nonrandom selection into the study, thus creating generalizability issues due to sample selection. As noted above, two of the three sections had a 100% participation rate, while the other had an 81% rate; thus we do not expect such selection to bias our results.

[29]We cannot use the number of silver bullets as a dependent variable for this analysis since no examiner found the silver bullet without expert advice, resulting in zero variation to explain in the outcome variable for this subsample.

8 | DISCUSSION AND CONCLUSION

Our study is motivated by potential biases that limit the effectiveness of ML process technologies and the scope for human capital to be complementary in reducing such biases. We interpret our results as follows. The promise of ML technology in the patent examination context lies in its ability to make superior predictions by identifying a narrower, more relevant distribution of prior art. However, when patent applications are characterized by (plausibly strategic) input incompleteness, the ML technology may be more likely to make biased predictions without domain-specific expertise. Additionally, vintage-specific skills help navigate the ML user interface. Taken together, our results speak to the larger debate about whether ML can replace each and every element of a managerial decision. Our findings related to the importance of domain- and vintage-specific human capital speak to the cross-sectional and time series dimension of the problem, and our observational and experimental analyses show where and why human capital complements ML.

Our study is subject to important limitations, some of which call for additional research. First, we focus on the early stages of the evolution of ML technologies in a one-shot experiment. This boundary condition is both a feature and a limitation. Initial conditions are important to study, as they have path dependent outcomes for the future, both in terms of adoption decisions, and the strategies which will enable better diffusion of adopted technologies. However, it is a limitation inasmuch as our experimental window did not enable us to examine prolonged association of workers with the technology, leaving us unable to examine performance improvements over longer periods of time. While the provision of expert advice and vintage-specific human capital increases initial productivity, it is unclear is whether constant exposure and learning-by-doing by workers would cause the relative differences between the groups to grow or shrink over time.

Second, our reliance on an experimental design and choice of MBA students as subjects was motivated by the need for a sample of highly skilled but heterogeneously specialized labor force of "novice users" of process technology.[30] However, as is true for all laboratory experiments, generalizability of the results is limited by the abstractions from reality, and applications to other relevant subpopulations of the labor force. While we expect our sample population of MBA students at a top tier business school to be similar to highly skilled individuals interested in intellectual property rights, the results of our experiments need to be replicated in similar contexts for confirmation. Also, we deliberately abstracted away from vertical differentiation (high vs. low skilled labor), but widespread use of new technologies may well require an inquiry expanding to this dimension too. Third, although our experimental sample is relatively large, finding the silver bullet was relatively uncommon. As a result, it is possible for that our results are driven by outliers. We hope that future research will investigate this issue in an even larger sample.

Finally, our research context and technology vintages are very specific—applicable to USPTO's development of the ML tool *Sigma* relative to Boolean search. ML technologies are themselves heterogeneous, as are the contexts in which they are deployed and the way they are deployed (i.e., with or without complex user interfaces). In particular, we note that Sigma is a relatively early stage ML tool, with perhaps less friendly user interfaces than might be ideal, and perhaps with less adversarial training than would be ideal. Accordingly, while our results

[30]This choice of experimental subjects permitted us to control for prior familiarity with both task and technology and isolate the effects of prior knowledge domains and domain expertise provision to randomly distributed subjects.

should be interpreted with caution, we urge scholars to add to the budding empirical research examining evolution in the productivity of all ML technologies, and their contingencies.

Limitations notwithstanding, our study makes several contributions to existing literature. We contribute to the emerging social sciences and management literature on bias in ML prediction by building upon and extending the adversarial ML literature in computer science. We shed light on salience bias arising from input-incompleteness as a distinct source beyond biases embedded in the training data or model algorithms. Such input incompleteness, particularly resulting from strategic action, may well characterize numerous business and socio-economic settings, and our study is the first to provide evidence on both its relevance for prediction, and offer a potential solution that does *not* require opacity of ML algorithms (Cowgill & Tucker, 2020).

Second, recent theoretical models and empirical findings related to AI and ML have stressed substitution effects of technologies on human capital, particularly when related to prediction (Agrawal et al., 2018; Autor, 2015; Benzell et al., 2015). Human capital in this literature stream is largely incorporated either in the form of vertical differentiation (high vs. low skilled workers), or due to comparative advantage in complementary tasks related to judgment, new discoveries, or tacit knowledge (Aghion et al., 2017; Agrawal et al., 2018; Autor, 2015). We extend this literature by recognizing that domain expertise can be complementary to ML algorithms when the input to the algorithm may be subject to input incompleteness.

Our results related to vintage-specific human capital also contribute to the literature on strategic renewal and the literature on career specialization (Becker & Murphy, 1992; Lazear, 2004; Teodoridis, 2017). The former acknowledges both imperatives and challenges of change (Agarwal & Helfat, 2009; Anderson & Tushman, 1990; Christensen, 1997), and in particular the extent to which complementarities with existing resources and capabilities determine the rate at which new technologies substitute for old technologies (Adner & Snow, 2010; Agarwal & Prasad, 1999; Bapna et al., 2013; Bloom, Sadun, & Van Reenen, 2012; Gambardella, Panico, & Valentini, 2015; Hatch & Dyer, 2004). Here, in addition to establishing the continued complementarity of domain expertise, we show ML technologies, at least in their incipiency, appear to privilege those with computer science and engineering backgrounds due to differential rates of absorptive capacity (Cohen & Levinthal, 1990). Such complementarities imply that firms should carefully consider adoption relative to the skillset of their workforce. For workers, the benefits and costs of career specialization must also be evaluated in the context of what knowledge and skills are foundational for complementarities with the latest technology vintages. An unresolved question is for *how long* ML technologies will privilege those with computer science and engineering backgrounds. If design choices reduce the need for specialized skills, then the ML-CS&E complementarity may be brief.

The study has several implications for practice. As applicable to USPTO, our study shows the ML technology does well for what it is programmed to do: find the most similar patents to the focal application. It may be the case that for most patent applications, this tool substantially reduces USPTO response times because examiners are able to more quickly reject claims. However, our study shows unsupervised examination by ML technologies or by novice examiners risks accepting patents based on the wrong set of prior art. Domain expertise is needed to improve the search strategy and a CS&E background is required to effectively operate the user-interface. Moreover, the USPTO is considering making the tool available to the public. While such a public move may prevent inventors with easily rejected patent applicants from applying in the first place (because they would know that it was not novel), it is also possible that it may increase the prevalence of strategic patent writing by allowing the applicants to see which prior

art the ML algorithm is bringing to the fore. In this case, domain expertise will be all the more important to identify these attempts.

9 | CONCLUSION

As is noted in Special Report by the Executive Office of the President (2016: 1), there is "great optimism about AI and ML in their potential to improve people's lives by helping solve some of the world's greatest challenges and inefficiencies." By studying relative productivity differentials between nascent ML and established technologies, we have highlighted the complementary role of two critical human capital attributes—domain-specific expertise and vintage specific skills. With these in conjunction, ML technologies may well be able to deliver on the optimism for the future.

ORCID

Evan Starr https://orcid.org/0000-0002-4368-1710

REFERENCES

Adner, R., & Kapoor, R. (2016). Innovation ecosystems and the pace of substitution: Re-examining technology S-curves. *Strategic Management Journal*, *37*(4), 625–648.

Adner, R., & Snow, D. (2010). Old technology responses to new technology threats: Demand heterogeneity and technology retreats. *Industrial and Corporate Change*, *19*(5), 1655–1675.

Agarwal, R., & Helfat, C. E. (2009). Strategic renewal of organizations. *Organization Science*, *20*(2), 281–293.

Agarwal, R., & Prasad, J. (1999). Are individual differences germane to the acceptance of new information technologies? *Decision Sciences*, *30*(2), 361–391.

Aghion, P., Bergeaud, A., Blundell, R., & Griffith, R. (2017). *Innovation, firms and wage inequality*. Retrieved from: https://scholar.harvard.edu/files/aghion/files/innovations_firms_and_wage.pdf

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction, judgement, and uncertainty* (No. 24243). NBER Working Paper.

Akerlof, G. (1982). Labor contracts as partial gift exchange. *The Quarterly Journal of Economics*, *97*(4), 543–569.

Alcacer, J., & Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics*, *88*(4), 774–779.

Allison, J. R., & Lemley, M. A. (2000). Who's patenting what-an empirical exploration of patent prosecution. *Vanderbilt Law Review*, *53*, 2099.

Anderson, P., & Tushman, M. L. (1990). Technological discontinuities and dominant designs: A cyclical model of technological change. *Administrative Science Quarterly*, *35*(4), 604–633.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias. ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Autor, D. (2014). *Polanyi's paradox and the shape of employment growth* (No. 20485). National Bureau of Economic Research.

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives*, *29*(3), 3–30.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, *118*(4), 1279–1333.

Baer, T., & Kamalnath, V. (2017). *Controlling machine-learning algorithms and their biases*. McKinsey & Company Website. https://www.mckinsey.com/business-functions/risk/our-insights/controlling-machine-learning-algorithms-and-their-biases

Bapna, R., Langer, N., Mehra, A., Gopal, R., & Gupta, A. (2013). Human capital investments and employee performance: An analysis of IT services industry. *Management Science*, *59*(3), 641–658.

Baracaldo, N., Chen, B., Ludwig, H., Safavi, A., & Zhang, R. (2018, July). Detecting poisoning attacks on machine learning in IoT environments. In *2018 IEEE international congress on internet of things* (pp. 57–64).

Bartel, A. P. (1994). Productivity gains from the implementation of employee training programs. *Industrial Relations: A Journal of Economy and Society*, *33*(4), 411–425.

Bashir, S., & Rauber, A. (2010). Improving retrievability of patents in prior-artsearch. In *European Conference on Information Retrieval* (pp. 457–470). Berlin, Heidelberg: Springer.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of Political Economy*, *70*, 9–49.

Becker, G. S., & Murphy, K. M. (1992). The division of labor, coordination costs, and knowledge. *The Quarterly Journal of Economics*, *107*(4), 1137–1160.

Benzell, S. G., Kotlikoff, L. J., LaGarda, G., & Sachs, J. D. (2015). *Robots are us: Some economics of human replacement* (No. w20941). National Bureau of Economic Research.

Bessen, J. E. (2016). How computer automation affects occupations: Technology, jobs, and skills. *Boston University School of Law*, 15–49. https://scholarship.law.bu.edu/faculty_scholarship/813

Black, S. E., & Lynch, L. M. (2001). How to compete: The impact of workplace practices and information technology on productivity. *The Review of Economics and Statistics*, *83*(3), 434–445.

Bloom, N., Sadun, R., & Van Reenen, J. (2012). Americans do IT better: US multinationals and the productivity miracle. *The American Economic Review*, *102*(1), 167–201.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)* (pp. 4356–4364). Red Hook, NY: Curran Associates Inc.

Brynjolfsson, E., & McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. *WW Norton & Company*, *2014*.

Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlström, P., ... Trench, M. (2017). *Artificial intelligence–The next digital frontier*. McKinsey Glob Institute, 47, 3–6. Retrieved from https://www.mckinsey.de/files/170620_studie_ai.ePdf

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 1–15.

Castanias, R. P., & Helfat, C. E. (1991). Managerial resources and rents. *Journal of Management*, *17*(1), 155–171.

Chari, V. V., & Hopenhayn, H. (1991). Vintage human capital, growth, and the diffusion of new technology. *Journal of Political Economy*, *99*(6), 1142–1165.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

Choudhury, P., Khanna, T., & Mehta, S. (2017). *Future of patent examination at the USPTO Harvard Business School case study* (pp. 617–027). Boston, MA: Harvard Business School.

Christensen, C. (1997). Patterns in the evolution of product competition. *European Management Journal*, *15*(2), 117–127.

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, *35*(1), 128–152.

Cowgill, B. (2017). *The labor market effects of hiring through machine learning working paper*.

Cowgill, B., & Tucker, C. E. (2020). Economics, fairness and algorithmic bias. *Journal of Economic Perspectives. SSRN* (in press). http://dx.doi.org/10.2139/ssrn.3361280

Crouch, D. (2014). *USPTO's swelling examiner rolls*. Retrieved from https://patentlyo.com/patent/2014/11/usptos-swelling-examiner.html

D'hondt, E. (2009). Lexical issues of a syntactic approach to interactive patent retrieval. In *The proceedings of the 3rd BCSIRSG symposium on future directions in information access* (pp. 102–109). https://dl.acm.org/doi/10.5555/2227296.2227313

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dickinson, Q. T. (2000). Remarks of q. todd Dickinson at pto day annual conference on patent and trademark office law and practice. *Journal of the Patent and Trademark Office Society*, *82*(3), 219–231.

Edmondson, A. C., Winslow, A. B., Bohmer, R. M., & Pisano, G. P. (2003). Learning how and learning what: Effects of tacit and codified knowledge on performance improvement following technology adoption. *Decision Sciences*, *34*(2), 197–224.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the Acquisition of Expert Performance. *Psychological Review*, *100*(3), 363.

Executive Office of the President. (2016). *Preparing for the future of artificial intelligence*. Washington, D.C.: Executive Office of the President.

Frank, M., Autor, D., Bessen, J., Brynjolfsson, E., Cebrian, M., Deming, D., ... Rahwayn, Y. (2019). Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, *116*(14), 6531–6539. https://doi.org/10.1073/pnas.1900949116

Gambardella, A., Panico, C., & Valentini, G. (2015). Strategic incentives to human capital. *Strategic Management Journal*, *36*(1), 37–52.

Gibbons, R., & Waldman, M. (2004). Task-specific human capital. *American Economic Review*, *94*(2), 203–207.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Government Accountability Office. (2016). *Intellectual property: patent office should strengthen search capabilities and better monitor examiner's work*. Report to the Chairman, Committee on the Judiciary, House of Representatives.

Greenwood, B., Agarwal R., Agarwal R., & Gopal A. (2019). The role of individual and organizational expertise in the adoption of new practices. *Organization Science*, *30*(1), 191–213.

Hatch, N. W., & Dyer, J. H. (2004). Human capital and learning as a source of sustainable competitive advantage. *Strategic Management Journal*, *25*(12), 1155–1178.

Hounshell, D. (1985). *From the American system to mass production, 1800–1932: The development of manufacturing technology in the United States* (p. 4). Baltimore, Maryland: JHU Press.

Jovanovic, B. (2009). The technology cycle and inequality. *The Review of Economic Studies*, *76*(2), 707–729.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293.

Knight, W. (2019). The Apple card does not "see" gender—And that's the problem. *Wired*. Retrieved from https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/

Kong, E. B., & Dietterich, T. G. (1995). Error-correcting output coding corrects bias and variance. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*, July 1995 (pp. 313–321). https://dl.acm.org/doi/10.5555/3091622.3091661

Krishna, A. M., Feldman, B., Wolf, J., Gabel, G., Beliveau, S., & Beach, T. (2016). User interface for customizing patents search: An exploratory study. In C. Stephanidis (Eds.), *HCI International 2016 – Posters' Extended Abstracts. HCI 2016. Communications in Computer and Information Science* (Vol. 617). Cham: Springer.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 275–284). ACM. https:/doi.org/10.1145/3097983.3098066

Lampe, R. (2012). Strategic citation. *Review of Economics and Statistics*, *94*(1), 320–333.

Lazear, E. P. (2004). Balanced skills and entrepreneurship. *The American Economic Review*, *94*(2), 208–211.

Lemley, M. A., & Sampat, B. (2012). Examiner characteristics and patent office outcomes. *Review of Economics and Statistics*, *94*(3), 817–827.

Merriam-Webster. (2018). Search for "Artificial Intelligence".

Mokyr, J. (1990). *Twenty five centuries of technological change: An historical survey* (Vol. 35). Abingdon, Oxon: Taylor & Francis.

Osoba, O. A., & Welser, W., IV. (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Santa Monica, California: Rand Corporation.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security* (pp. 506–519). https://doi.org/10.1145/3052973.3053009

Polonski, V. (2018). AI is convicting criminals and determining jail time, but is it fair? *World Economic Forum*. Retrieved from https://www.weforum.org/agenda/2018/11/algorithms-court-criminals-jail-time-fair/

Robinson, D., & Koepke, L. (2016). Stuck in a pattern: Early evidence on 'predictive policing' and civil rights. *Upturn*. Retrieved from http://www.stuckinapattern.org

Simon, H. (1991). Bounded rationality and organizational learning. *Organization Science*, *2*(1), 125–134.

Sweeney, L. (2013). Discrimination in online ad delivery. *arXiv preprint arXiv:1301.6822*.

Teodoridis, F. (2017). Understanding team knowledge production: The interrelated roles of technology and expertise. *Management Science*, *64*(8), 3625–3648.

Thoma, G., Torrisi, S., Gambardella, A., Guellec, D., Hall, B. H., & Harhoff, D. (2010). *Harmonizing and combining large datasets—An application to firm-level patent and accounting data* (No. w15851). National Bureau of Economic Research.

Thornton, R. A., & Thompson, P. (2001). Learning from experience and learning from others: An exploration of learning and spillovers in wartime shipbuilding. *American Economic Review*, *91*(5), 1350–1368.

Turk, S. A. (2006). The proper method for using dictionaries to construe patent claims. *Chicago-Kent Journal of Intellectual Property*, *6*, 43–65.

Verberne, S., D'hondt, E. K. L., Oostdijk, N. H. J., & Koster, C. H. (2010). Quantifying the challenges in parsing patent claims. In *Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval at ECIR* (pp. 14–21). https://repository.ubn.ru.nl/handle/2066/84168

How to cite this article: Choudhury P, Starr E, Agarwal R. Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strat Mgmt J*. 2020;41:1381–1411. https://doi.org/10.1002/smj.3152