AI and the Creation of Knowledge Gaps: The ethics of AI transparency

Kirsten Martin
University of Notre Dame

Bidhan Parmar
University of Virginia

ABSTRACT

Managers have stakeholder obligations to justify their decisions align with the values and norms of the organization, further the mission of the firm, avoid breaking relevant laws and regulations, and promote the long-term interests of the firm. These obligations do not disappear when firms adopt AI decision systems. We introduce the concept of the AI knowledge gap – where AI is designed to supply limited information about its operations that precludes managers from meeting stakeholder demands for information justifying firm decisions. We outline the types of knowledge required to ensure the manager and firm can meet their existing obligations to stakeholders. Given these existing obligations, adopting recommendations from a 'black box' AI program, where the manager cannot justify the firms' decisions to stakeholders, is unethical. We argue that adequate knowledge about an AI model is not a negotiable design feature but a strategic and moral requirement.

Keywords: AI, Transparency, Inscrutable AI, Explainable AI, AI Ethics

AI and the Creation of Knowledge Gaps: The ethics of AI transparency

Organizations are adopting AI with a fever pitch. The market for AI is estimated to reach $738B by 2030, and manufacturing, healthcare and finance and the largest industries by market share.[1] Firms have ventured further into opaque algorithmic decision-making (ADM) systems, are difficult for humans to interpret, and result in corresponding calls for greater algorithmic transparency (Felzmann et al. 2020; Mökander et al. 2021). The complexity of machine learning models, the lack of tools to visualize data and outcomes, technical illiteracy, and corporate secrecy contribute to AI decisions becoming increasingly opaque to managers (Burrell 2016; Diakopoulos 2020). This opacity enables strategic and ethical missteps, such as unfair and discriminatory predictions and the dehumanizing treatment of individuals, while simultaneously making it more difficult to define clear accountability for such problems.

The holy grail for algorithmic decision-making systems, it seems, would be AI models that have the necessary complexity to be helpful in ambiguous decision contexts with the corresponding clarity needed to fulfill existing business obligations. Firms have obligations and are accountable to many stakeholders, requiring organizational actors from top management teams to line operators to justify their decisions. For example, firms have obligations to adhere to state, federal, and international laws, minimize harm to employees, make choices that do not erode the brand and mission of the organization, and preserve and strengthen stakeholder relationships. Organizations implement AI models to identify applicants to hire and employees to fire, to price products automatically, and to recommend courses of action from medical diagnoses to parole terms. And organizations have obligations to justify these decisions and answer to stakeholders (Tigard 2021) *regardless* of how AI augments the decision. Despite these responsibilities, calls for more transparency of AI models have largely ignored the adopting organization as a stakeholder with specific ethical concerns (Buhmann and Fieseler 2021) and within a particulat power structure (Bleher and Braun 2023).

Within the current conversation about the responsible use of AI, scholars can mistakenly assume that organizations are primarily concerned with avoiding mistakes and increasing acceptance. This impoverished view of organizational motives leads to expectations that organizations will accommodate the new AI technology however it is designed – particularly

[1] https://www.statista.com/outlook/tmo/artificial-intelligence/worldwide#market-size

within management (Asatiani et al. 2020; Benbya et al. 2020). Significantly less attention has been paid to examining the adopting firm's related strategic and ethical obligations and how to put ethics into practice with the implementation of AI in organzations (Bleher and Braun 2023). Specifically, what obligations do managers and firms have to their current stakeholders, and how will adopting an AI model impact those obligations?

The goal of this paper is to outline the current obligations of firms and how those obligations should guide the adoption of AI models. Firms and managers have existing obligations to their stakeholders, including a demands for reasons for the firm's actions and decisions (Coeckelbergh 2020), and we argue that these obligations constitute a type of 'demand for knowledge' about AI models so that managers and executives can continue to justify their business decisions to their stakeholders. We argue that the current use of AI puts managers and firms in a risky knowledge gap. Where there is an increased demand to justify AI-augmented decisions to stakeholders and a corresponding shortage of information for managers to offer in response to those same ADM systems.

This paper highlights a firm's existing strategic and moral obligations as constituting design requirements for developing and adopting AI models. First, we re-center the most powerful stakeholder of the AI model – the firm that purchases and implements the model – to take responsibility for their choice of a particular AI model and how they use it. Second, by making the moral obligations of business a key input into knowledge requirements about AI into design criteria or external constraints on computer or data scientists, we extend AI governance practices into the procurement process, not just the development process (Dor and Coglianese 2021). This demand for particular types of information from adopting firms should shape the design and development of ADM systems. In this way, we return to the norms of engineering to solve problems within external constraints: the adopting firm's existing obligations should be the external constraints on the design and devleopment of AI.[2]

Importantly, an AI program designed as a 'black box' would preclude a firm from meeting existing obligations. According to this argument, adequate comprehension and knowledge about an AI model is not a negotiable design feature but a strategic and moral requirement. In other words using black box AI models within algorithmic-decision systems is

[2] Engineering works within constraints, by definition, and ethics and values can be a part of that constraint in the design of technology (Nair and Bulleit 2018).

unethical when it prevents managers, executives, and organizations from meeting their existing moral obligations to stakeholders. Current research and guidance take AI as a given and requests *more* transparency and knowledge as may be technologically possible. Our argument flips the requirements regarding accuracy, efficiency, and transparency. The (mistaken) assumption is that to benefit from the claimed accuracy and efficiency of AI, managers must settle for as much transparency as they can from the model. In contrast, we argue that an ethical organization cannot meet its obligations to justify its decisions without adequate knowledge about algorithmic decision-making. If AI is designed in ways that make it inscrutable, the firm that *develops* the AI program would need to take responsibility for how the program performs, who is negatively impacted by it, and any value that is lost by using the AI program. Instead, these developing firms want to be treated as if they are an authority but not take on the associated responsibility of being an authority.

SUPPLY OF KNOWLEDGE: TRANSPARENCY, EXPLAINABILITY, & INTERPRETABILITY.

Organizations employ AI models as part of a larger algorithmic decision-making (ADM) system to make business decisions - consisting of both humans and technology (Diakopoulos 2020). The AI model, when implemented, takes over tasks previously completed by employees and less-developed technology (Martin 2019). For example, an AI model that scans and categorizes job applicants' resumes takes the place of the employees who once read those resumes and made employment judgments. However, these AI models can be challenging to understand for several reasons. Models are purposefully developed to be secretive and proprietary, and models may be designed to require specialized knowledge not currently widely available within organizations and stakeholder groups (Burrell 2016; Pasquale 2015; Selbst and Barocas 2018). The opaqueness of these AI models becomes problematic when the larger ADM system and the organizations they are a part of must justify their decisions to others. Yet, the obscurity of the AI model hinders the ability of the larger ADM system to justify outcomes and the organization to understand how to be accountable for those outcomes. For example, when employee groups ask why minority applicants are less likely to be hired after an AI model is implemented, HR managers do not know how to respond because they do not understand how the AI model works.

Types of Knowledge Supplied

The current approach to this lack of understanding about ADM systems is to focus on the supply of more information about how the AI model works so that managers can be held accountable for the adverse decisions, predictions, and outcomes of algorithmic-decision systems (Ananny and Crawford 2018; Diakopoulos 2016; Kroll et al. 2017). Three concepts have been used to explore the supply of knowledge about algorithmic decision systems: transparency, explainability, and interpretability. Many times these concepts are used interchangeably. For example, the desire for transparency is sometimes framed as requiring an explanation that is interpretable. Here, however, we explain how the three main approaches have all been developed to increase the available knowledge about AI models.

AI transparency aims to provide enough information so that others can understand the performance of the AI model, thus making the ADM system knowable (Diakopoulos 2020; Rader et al. 2018). However, the definition of transparency varies, and the specific nature of transparency changes based not only on the person receiving the information but also based on *why* the person needs the information (Lipton 2018). For example, transparency may be useful to provide notice of the mere use of AI for user consent or in the service of complying with GDPR (Edwards and Veale 2017; Felzmann et al. 2020; T. W. Kim and Routledge 2020; Selbst and Barocas 2018). The strongest version of transparency requires that the model's functionality can be comprehended entirely by an individual (Mittelstadt et al. 2019). Yet, the majority of the calls for transparency cast a much narrower goal for information about the AI model; specifically, to answer questions about a specific purpose from an audience and not to require the entirety of the functionality to be comprehended by a person.

Transparency is a broad term covering the communication of information about an AI model includes the more specific concepts of explainability and interpretability. ***Explainable AI (XAI)*** refers to the suite of techniques created to make a given AI model better understood (Felzmann et al. 2020; Speith 2022). For example, in Figure 1, XAI is when a second (post hoc) model is created to explain the first (black box) model (Rudin 2019). As such, XAI models more closely resemble summary statistics of the model rather than actual explanations (Rudin 2019). Research in explainable AI investigates what types of explanations are possible, what constitutes an explanation, and how to make people understand the explanations provided (Páez 2019).

A critical assumption in the explainable AI approach is that AI opacity is a given attribute of an AI model. And the goal is to try to get closer to explaining the micro-level relationship between the source data and a prediction or outcome. Such an approach can be limiting, as Christina Rudin notes, since maintaining the black box and offering an 'add-on' XAI model to explain the black box model allows the developing firm to continue to benefit from the black box model obscuring the problematic results and design (Rudin 2019).

A second technical solution for greater transparency is to create ***interpretable*** models rather than "trying to explain black box models" (Rudin 2019). Where explainability is an intermediate interface between a model and humans and takes the opaqueness of the AI model as a given, interpretability is an attribute of the AI model and is an attempt to design the original AI model in a way that is more easily understood (Arrieta et al. 2020). Rudin notes that interpretable systems take more effort and domain expertise to construct (Rudin 2019), and the need for interpretability arises when the AI model as designed does not work as expected (Lipton 2016).

Figure 1: Supply-Focused Knowledge About AI Models

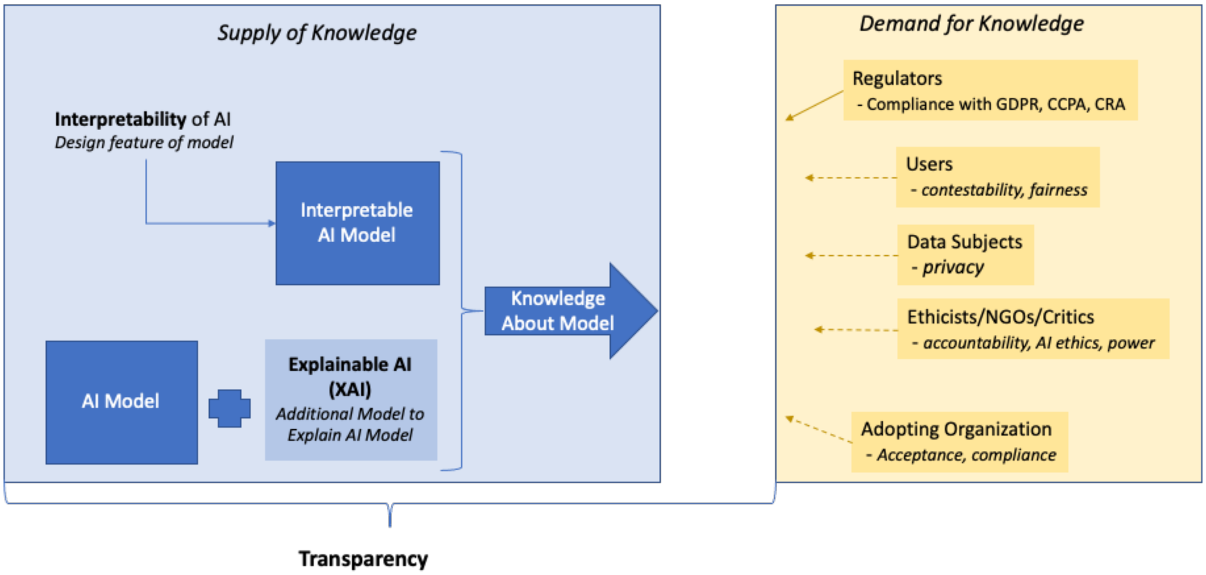


Figure 1 illustrates the three main approaches to producing more knowledge about an AI model. The main audiences are assumed to be independent stakeholders interested in what data is used, how the model is developed, and when and how the model is implemented. The primary audiences mentioned in the design and testing of explainable and interpretable AI scholarship are regulators, critics generally, 'society,' and possibly the subjects of the model. Of note to us is that the firm adopting the model for a specific decision context is often not included as an audience to consider, or the firm is framed as having simple requirements to (1) have the model accepted by employees and (2) avoid risks (Felzmann et al. 2020). Regulatory bodies are the most consistent audience with a direct relationship with the developing firm.

Problems with Focusing Only on Supply of Knowledge

The current approach relying on the production of knowledge about AI models has introduced ethical issues to developers with the hope that firms that develop AI models will think before producing opaque models. However, the reliance on production, rather than demand, for AI knowledge has several weaknesses: being limited by technical capabilities and being focused on only some audiences.

First, the production of knowledge is limited by that which is technically possible or imagined by developers to offer *more* explainable models while maintaining the claimed 'accuracy'(Arrieta et al. 2020). As such, knowledge about the AI model produced by the developing firm "might be strategically shaped, distorted, or unreliable and therefore less conducive to accountability" (Diakopoulos 2020). Developing firms produce knowledge based on what they can currently provide and what they want to provide; thus, the knowledge produced may or may not meet the needs of inquiring stakeholder groups. This approach takes the opacity of AI models as a given and treats any additional information as 'nice to haves,' rather than 'must-haves.'

In addition, by focusing on technical solutions that better explain how inputs are translated to outcomes, critics of AI transparency argue that providing additional information about an AI model generally introduces risk to the developing firm in terms of giving away secrets and providing information so that the model can be gamed (Burt 2019; Tsamados et al. 2021). In allowing the developing firm to decide how to explain their model and how understandable to make it, the motivation for firms to provide selective statistics about their model or keep the model proprietary still exists in the market (Martin 2023).

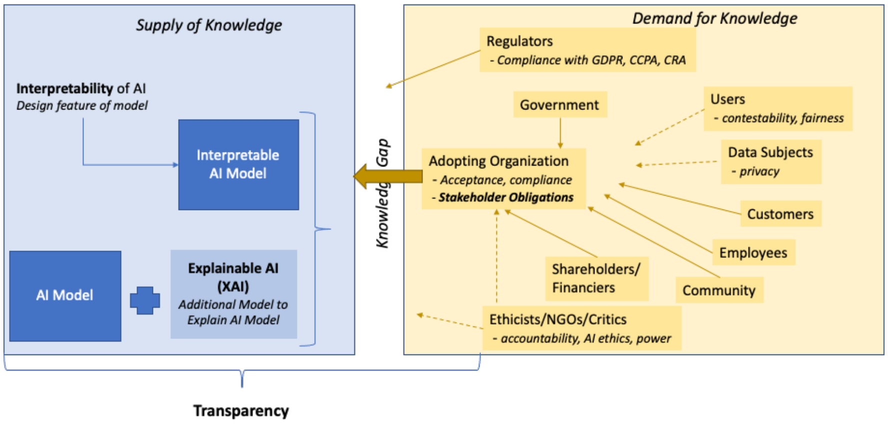
Second, who to consider when supplying more knowledge about an AI model frequently does not include the requirements of the more powerful market actor: the adopting firm. The focus on supplying more technical knowledge about an AI model only recently considered audiences when producing knowledge (Arrieta et al. 2020; Diakopoulos 2020; Páez 2019). And, the main audiences are often the computer and data scientists themselves (Arrieta et al. 2020; Páez 2019). Developers are encouraged to consider the audiences' voices and concerns when creating knowledge about an AI model. Yet, in practice, these stakeholders are not seen as creating a binding requirement for information, resulting in a situation where some stakeholder concerns can be considered secondary. Notably, most of these audiences in Figure 1 have no relationship with the developing firm. Thus far, strong demands for knowledge from AI models are limited to only freedom of information requests (Diakopoulos 2020) and compliance with some laws, such as GDPR, that directly impose costs on the organization developing the AI model (Felzmann et al. 2020). However, this ignores the current interests, concerns, and obligations organizations have to justify their decisions. For example, an HR manager must justify their application filtering results so they can build trust with the employees and applicants and meet their obligations to supervisors, top management teams, financiers, employees, and other stakeholders. Organizations purchasing and adopting AI models are assumed only to worry about compliance and avoiding mistakes and not about meeting their current obligations (Langer et al. 2021; Someh et al. 2022).

As Selbst and Barocas note, asking, 'what types of explanations are possible,' and 'what types of explanations are most useful,' are the wrong questions (Selbst and Barocas 2018). Selbst and Barocas focus on *legal* constraints such as due process and substantive laws, such as, GDPR and credit reporting, that directly impact AI model developers. Our contribution to this line of thinking is to argue that firms have strong interests and obligations within current stakeholder relationships beyond GDPR that regularly become purchasing requirements and preferences to operationalize. For example, an aerospace company would put substantial design requirements within a contract for steel alloys. Or a firm would put performance requirements in the outsourcing agreement for customer service. Typically, firms purchasing new technology incorporate their obligations to stakeholders in the purchase agreements and design criteria.

Figure 2 illustrates the more contextualized stakeholder relationships that drive firm obligations. Firms regularly are accountable to customers, shareholders, employees, communities, suppliers, users, and state and federal regulators. Firms are accountable for both the decision to adopt an AI model within an ADM system and the ongoing decisions made within the ADM system. This figure also re-centers a powerful stakeholder – the adopting organization – who purchases and adopts the AI model for use within a larger ADM system.

Where previously adopting firms were assumed to only be concerned with general risks and acceptance, if at all, here we acknowledge the adopting firm is an actor with the most power over the developing organization in the marketplace. If firms decide not to buy an AI solution because it does not meet their requirements, firms developing AI will take notice. We now explore the adopting firms' existing stakeholder obligations that should drive design and purchasing requirements for AI models.

Figure 2: Stakeholder-Focused Demand for Knowledge about AI Models.



DEMAND FOR KNOWLEDGE: FIRM OBLIGATIONS TO STAKEHOLDERS

Firms have multiple obligations to their stakeholders. For example, one obligation is to pursue strategies that are in the company's long-term interest (Blair and Stout 2001; Stout 2012). This can be through focusing on specific stakeholders or, as recommended in both strategy and business ethics, creating value for internal and external stakeholders (Freeman 1984; Freeman and Phillips 2002). Companies develop a mission, vision, norms, and values to guide executives and managers throughout the firm to create a consistent set of outcomes. These various obligations can be seen as grounded in different ethical traditions – for example, laws about protecting the rights of employees are grounded in deontological traditions that focus on explicit norms, rules, and laws. Obligations that minimize harm on customers can have their roots in consequentialist traditions that generally emphasize creating the greatest good for the greatest number. Similarly, obligations to protect and promote the firm's reputation and character is connected to traditions of virtue ethics, and finally obligations to look after the welfare of key stakeholders can be seen to derive from an ethics of care. Therefore, managers find themselves in complex moral territory, with multiple (sometimes competing) moral obligations that derived from different perspectives and traditions. They also interact with stakeholders who rely on different traditions to make sense of managerial choices. Not meeting those obligations is not only bad business strategically but will be seen as unethical by those stakeholders who expect the firm to honor those obligations. Thus managers must be able to identify various moral obligations and work with stakeholders to define and deliver on shared expectations.

Managers and executives make decisions knowing the current requirements, goals, values, norms, and obligations of the firm: both the obligations to act within the interest of the firm and be able to show that all decisions meet that standard.

Therefore, firms and managers have obligations to explain and justify decisions to stakeholders to show that the actions taken are compatible with the values and norms of the organization, further the mission of the firm, avoid breaking relevant laws and regulations, and promote the long-term interests of the firm. Explaining and justifying decisions increases perceptions of fairness, legitimacy, trust, and predictability with these stakeholder groups.

For a given action or decision, managers have stakeholders due to *who they are* (which firm) as well as *what they are doing* (the decision context). For example, a bank executive making a hiring decision has specific stakeholder obligations both due to the structure of their industry (competitors, banking regulators, suppliers, customers) as well as due to the decision (applicants, employees, employment regulators, etc).

Justifications to stakeholders

Organizations intentionally or unintentionally impact their stakeholders in many decisions, such as closing a production plant, firing an employee, dumping chemicals in a river, changing a marketing channel, releasing a new product, and adopting new technology. And organizations need to be able to provide reasons and a rationale for their actions. When they cannot, stakeholders will try to hold the firm or individual managers responsible for a lack of justification through the market or even escalate their grievances through the courts, which will ultimately hold the firm accountable and demand reasons

If an employee at a local manufacturing plant arrives at work one day, only to hear that the company has closed the plant, their trust and willingness to work with the company are radically diminished if the company gives unacceptable reasons like, “We just thought we’d have more fun in a new location.” The reasons we give each other are critical to building predictability and trust with our stakeholders. By sharing the reasons for our actions, we can better gauge whether a party is trustworthy and our interactions should continue, or if we see the world differently and we should end our coordination (P. H. Kim et al. 2006).

Managers have obligations to justify decisions within their span of control to their stakeholders to show that the actions taken are compatible with the values and norms of the organization, further the mission of the firm, avoid breaking relevant laws and regulations, and promote the long-term interests of the firm. Managers’ obligations to justify decisions to stakeholders creates a demand for knowledge that the manager must meet. Managers need requisite knowledge to meet this demand. Breaking these obligations will be seen by stakeholders as unethical because not only could this mean the manger is not acting in the long term interest of the firm, and therefore not doing the job they are paid to perform, but not meeting obligations in general is considered unethical since the manger is paid not only to create value for the firm and stakeholders but also meet their obligations that are in the best interest of the firm.

Since a firm has obligations to justify decisions to stakeholders, then when deciding to augment a decision, such as who is committing fraud, who to hire, or who to promote, with an AI model, managers need the requisite knowledge to meet these obligations. In other words, stakeholder obligations define the demand for knowledge about their firm's decisions.

KNOWLEDGE GAP

The use of AI does not change the reason-giving preconditions, obligations, and practices that come with being a manger or executive in a company. Still, AI models can challenge the ability to provide those reasons.[3] When managers treat AI as a black box, implement AI with little human intervention, and allow models to learn from data without supervision, firms can find themselves stuck between having to give justifications to their stakeholders and not having enough information, resulting in diminished trust and relationships. Such managers are in a knowledge gap: where the demand for the knowledge needed to meet the obligations to stakeholders exceeds the supply of knowledge available with the introduction of an opaque AI model.

This issue of a knowledge gap is not new for firms. When firms started outsourcing production and extending their supply chain to new countries, firms did not initially know enough about the working conditions of their partners while still being held responsible for their supply chains' actions. The demand for responsible supply chains (from customers, governments, activist groups, and firms themselves) has led supply chain managers to vet their suppliers further, ensuring fair labor practices and avoiding illegal practices like child labor, unsafe working conditions, sustainability, and fair wages. Figure 3 illustrates the knowledge gap and the state of equilibrium is designated by the dotted line – where a firm can meet the knowledge demands of its stakeholders. Innovations in supply chains, such as global sourcing, initially created a temporary knowledge gap whereby the demand for knowledge by stakeholders remained the same, but the supply of knowledge as to the workings of the supply chain was diminished. In Figure 3, the introduction of the global supply chain decreased the supply and moved the point to the left and into a state of disequilibrium. (A1→A2).

[3] Ironically, AI can be used to better understand how and why decisions have been made in the past – such as insurance claim adjudication, mortgage lending, hiring, etc. AI need not always limit the supply of knowledge an can provide insights into manager decisions if designed with this as a goal.

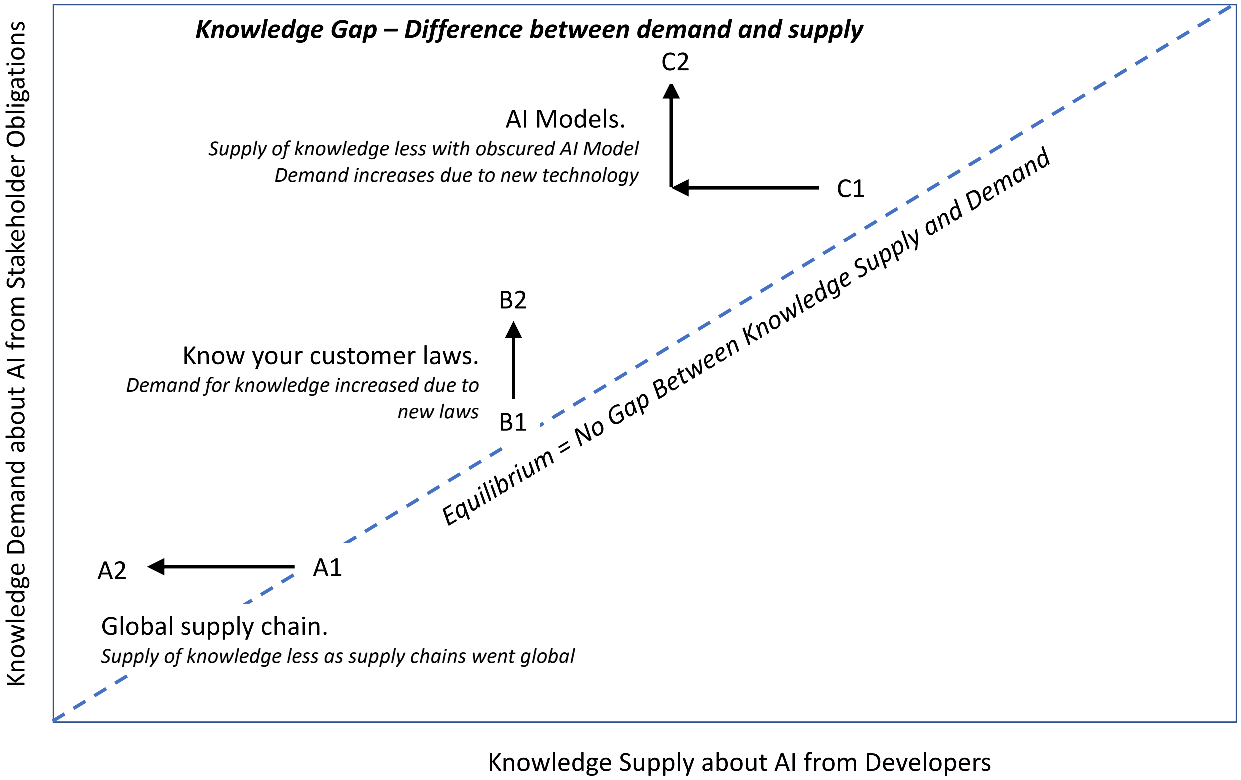
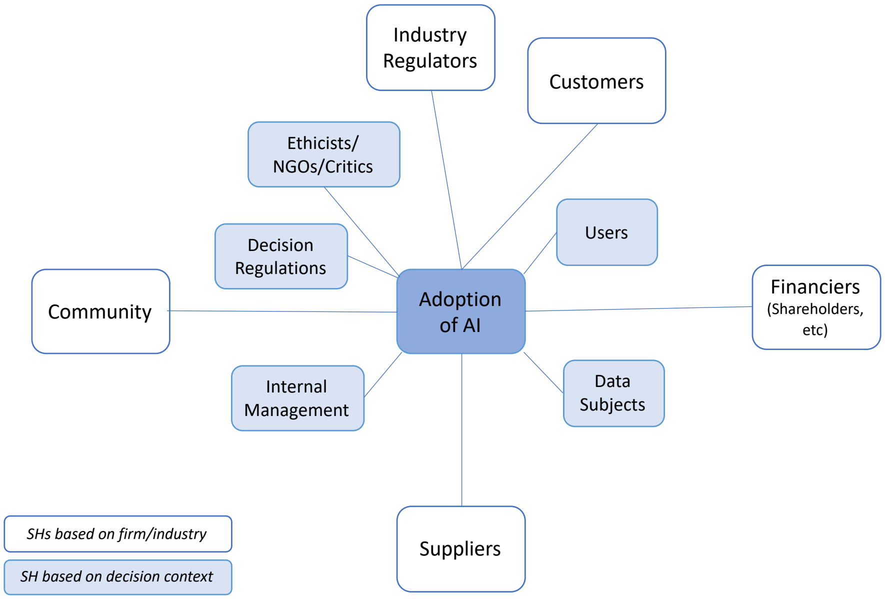

Knowledge Gap – Difference between demand and supply
C2
AI Models.
Supply of knowledge less with obscured AI Model
Demand increases due to new technology
C1
Equilibrium = No Gap Between Knowledge Supply and Demand
B2
Know your customer laws.
Demand for knowledge increased due to new laws
B1
A2
A1
Global supply chain.
Supply of knowledge less as supply chains went global
Knowledge Demand about AI from Stakeholder Obligations
Knowledge Supply about AI from Developers

Figure 3: Drivers of the AI Knowledge Gap

At other times the knowledge gap can be increased by new demands such as new regulations, standards, or norms. Regulations such as "know your customer" laws for banks created a need for banks to capture and report information about their customer base to avoid dealing with terrorist groups, money launderers, and other illegal activities based on the firm's industry (B1→ B2). Changing societal norms around advancing DE&I initiatives and protecting the rights of LGBTQ+ employees have created new obligations and demands for knowledge where firms are being asked to report information and explain how their employment decisions meet these obligations. Increased demand for knowledge by stakeholder groups led to firms capturing, analyzing, and reporting more data. In Figure 3, the demand for more knowledge created by new norms and laws shifted the firm up the y-axis and into a state of disequilibrium; firms adapted by creating a greater supply of knowledge to meet that demand.

Importantly, the demand for knowledge can be from stakeholders based on the firm/industry as well as the decision context. For example, the adoption of robotics within manufacturing did not absolve the firm of their obligations to ensure their products were created to meet industry quality and safety standards. Executives also had to ensure that the adoption of robotics in manufacturing was consistent with the firm's mission, values, and priorities an met employment safety requirements. Instead, these business requirements and obligations became design criteria for those developing and manufacturing robots. For example, OSHA has guidelines for the safe deployment of robots in auto manufacturing that companies must comply with – since the 1980s.[4]

In summary, managers and firms may find themselves stuck in a knowledge gap when adopting an AI model. AI can be designed in opaque and inscrutable ways, supplying little knowledge about how and why decisions are made. On the other hand, the demand for knowledge from stakeholders does not disappear. Indeed, demand may even increase as the use of AI increases the scale and speed of impacts on stakeholders. So, managers may have less knowledge to answer an increasing number of questions to fulfill their obligations to internal and external stakeholders. This gap increases the likelihood that managers and firms may be unable to meet their obligations and lose trust and goodwill from their stakeholders (C1→ C2).



[4] For an example of the regulations that shape the safe and responsible use of robots in the workplace: https://www.osha.gov/enforcement/directives/std-01-12-002

WHAT FIRMS NEED TO KNOW TO ADOPT AI

What do managers need to know to adopt AI responsibly? Our goal is not to specify an exhaustive list of issues that will work for all managers in all companies, but to provide a starting set of questions that can help business leaders better articulate the complex stakeholder obligations in regards to their specific use of ADM (Gebru et al. 2018; Mitchell et al. 2019).

For a given implementation of ADM, a firm would need to first identify the stakeholders based on both (1) who they are (the firm/industry) as well as (2) what they are attempting to do (the decision context). The goal of this set of questions is to identify the multiple moral obligations across a variety of ethical traditions (Deontology, Consequentialism, Virtue Ethics, and the Ethics of Care), because managers have obligations that stem from each of these traditions and stakeholders will utilize these traditions in arguing for their perspective. However, the premise of the argument here is that these obligations are legitimate and that abiding by these stakeholder obligations is ethical and a part of any executive's job. In other words, breaking stakeholder obligations is unethical based on both a deontological argument (ethical acts are those that conform to moral obligations) as well as a social contract argument (executives willingly take jobs knowing the norms of that job as well as the negotiated obligations inherent to the position). In addition, the adherence to existing obligations is part of the job or role of the executive and adhering to those obligations would acting in furtherance of excellence or showing good judgment (virtues). Finally, treating stakeholders and their obligations with respect *even when doing so is not easy* is evidence of an ethics of care as to the situated concerns of those staekhodlers – particularly vulnerable stakeholers. Adherence to obligations is normally not justified by short-term, myopic consequentialism (a manager should act for the short term benefit of herself) but is normally justified by a long-term consquentialism focused on maximizing the long term value of the firm.

Stakeholder obligations and two cases are displayed in Table 1. The first case is based on an HR-related program that predicts 'successful' candidates to hire within the tech industry (Dastin 2018). This program was based on the previous decade of hiring at the company; the program predicted how the company should hire if they wanted to hire the same way they had in the past. Unfortunately, the program 'learned' from the training data that gender mattered more than competence based on how the company had been hiring. This lead to the program choosing applicants based on different proxies for gender and while ignoring whether or not the applicant was actually qualified.[5]

For the second case, an insurance company used a program to adjudicate health insurance claims (Rucker et al. 2023a). The company's doctors, responsible for reviewing the patient records, coverage policies, and applying their expertise to each case, could instantly reject claims on medical grounds within an average of 1.2 seconds by signing off on the program's decisions in batches (and without opening the patient file). "We literally click and submit," one former doctor said. "It takes all of 10 seconds to do 50 at a time"(Rucker et al. 2023a). A former company executive decribed the system as "built to deny claims." And while only 1 in 5 denials were appealed, about 80% of those appeals were successful. A congressional committee as well as state and federal regulators began to ask the firm to produce documents and explain the program (Rucker et al. 2023b).

Table 1 illustrates how the obligations to stakeholders can drive design requirements for specific ADM programs for these cases. Taking this approach allows for some ADM programs to have *more* obligations – based on the firm/industry or decision context or both – and some to have fewer obligations. For example, insurance is a heavily regulated industry where employment or hiring is a heavily regulated decision context. Other examples we explore in the implications could have both a less regulated industry (e.g., AI firms right now) or a less regulated decision context (e.g., signature identification). The questions should be further developed by the adopting firm to create design criteria for the development and deployment of an AI model.

[5] This was because the training data was labeled 'hire' and 'not hire' based on who was hired in the past at the company. Since those hired and not hired were *equally quafliied* the model did not 'learn' that qualifications such as programing language experience differentiated the candidates. The one attribute that did differentiate those hired versus not hired was gender.

TABLE 1 – Stakeholder Obligations Based on Firm/Industry.

Firm Stakeholder Obligations	Insurance Claim Example	HR Example
Top Management Team		
What are the goals of this firm, division, and unit, and how does the use of AI support those goals?	How does the claim adjudication ADM improve "health and vitality" per our goals and mission? What metrics are needed to show such an improvement?	How does this program help us achieve our employee goals across the firm in terms of competence, skill sets, and diversity of skills, background, etc?
Which values and norms do the AI reinforce, and which may it jeopardize?	How could the use of AI to deny clams jeopardize values like trust, fairness, and "caring deeply about our… patients"? How can we show the new program supports our values?	How does the program reinforce or undermine our values of treating our employees with dignity and respect or with fair procedures.
Board of Directors		
Does the use of AI potentially undermine any firm stakeholder's rights or our obligations to that stakeholder?	How does the use of AI to deny claims meet the rights of patients to have a medical director to "examine patient records, review coverage policies and use their expertise" per state laws?	Do we have existing obligations to suppliers or customers to report our hiring practices (e.g., federal contracts)?
Does the use of AI add any additional financial risk to the firm we need to disclose?	How many new lawsuits or regulatory investigations could we encounter from the use of the claim adjudication ADM system as designed?	How many new lawsuits or regulatory investigations could we encounter from the use of the HR ADM system as designed?
Firm Regulators		
Who are our state and federal regulators and what do we need to report? What are their concerns?	How does the program meet state regulators of health insurance requirements around the reasons why claims can be denied? Are illegtimate factors (likelihood to appeal, race, sexual orientation, gender, etc) being used to deny claims? How does the use of the program violate state laws around the rights of patients to a "thorough, fair and objective investigation"?	Do we have the required information for state and federal employment laws (e.g., Title VII, ERA)? Can we ensure that our employment decisions abide by these same employment laws?
External Stakeholders (customers, suppliers, community)		
What are the systemic/societal impacts?	How does the program mitigate discrimination against customers with medical conditions that are rare, complex, or underrepresented in the data. Potential to impact lower SES customers negatively if their claims are rejected more than other groups.	How does the model allow us to ensure we do not contribute to unequally distributing hiring and entrenching inequality within the tech industry?

Firm Stakeholder Obligations	**Insurance Claim Example**	**HR Example**
Users/Data Subjects		
What are the current valid criteria used to make this kind of recommendation?	How does the program make decisions based on only "members benefit plans and clinical criteria in compliance with state and federal laws": i.e. Insurance policy coverage, specific procedures, past medical history, physician's recommendation,	Can we ensure that valid hiring criteria are being used – education, coding experience – while making sure invalid criteria are not being used (race, gender). For Amazon, the program ignored valid criteria and use invalid criteria.
Are users usually able to question or contest decisions?	How does the use of this program in any way preclude or hinder the ability of patients to appeal their decision? Does the company have the required information to process the appeal within state insurance laws?	How do subjects – applicants, employees – have the ability to contest the decision of the ADM? What information would be needed for users to have the ability to contest?
Internal Stakeholders		
What guardrails are in place to prevent harm and any violation of rights?	How can we ensure we do not deny a patient who needs a lifesaving procedure and dies. While not done in this case, a doctor should review claims that are rejected and track the percentage of claims denied by protected class.	While not done in this case, constant monitoring of decisions should highlight whether the people chosen for hire are a fit with firm goals and ensure illegal criteria are not being used.
How will mistakes be identified, judged, and addressed?	While not done in this case, what information is needed to trach the percent of denials that were legitimate as well as the percent of approvals that were legitimate, based on follow up, investigations, and testing	What information is required to monitor who is hired and who is not to see trends over time? How will we be able to see if the ADM is making mistakes and if mistakes are fairly distributed?
Which stakeholders to this decision could be harmed and who could benefit?	How does this program add costs to the most vulnerable (patients in need of medical care) while benefiting the most powerful (insurance firm)?	How does this program add costs to the most vulnerable (potential employees) while benefiting the most powerful (tech firm)?
Who is accountable internally for this decision and what information do they need?	While not done in this case, an individual or department should be named who is responsible for ensuring the ADM for claims both to answer internal stakeholders and ensure any decision is contestable.	Who (individual or department) is responsible for ensuring the ADM for hiring both to answer internal stakeholders and ensure any decision is contestable.

Questions such as, 'what are the laws and regulations related to this decision,' and 'does the use of this AI technology undermine any stakeholder's rights' can help managers identify and understand their current obligations. For the HR case, HR managers must check current HR regulations and policies around acceptable promotion criteria. Additionally, thinking about using AI from the perspective of a potentially harmed stakeholder can ensure that less formal norms and obligations are also surfaced. Questions about who is responsible for an error and what information is needed to ensure mistakes will be proactively identified and addressed can enable firms to minimize reputational damage and harm before they get out of hand. In the HR case, for example, managers may commit to reviewing key statistics about the population of employees who were hired and those that were not, ensuring that there are only differences in acceptable categories (such as ability and performance) and not in unacceptable categories such as (gender, race, or sexual orientation). In addition, firms have obligations to subjects and users to allow HR decisions to be contested and would need the knowledge required to allow such appeals.

Together by identifying the relevant stakeholders and the firm's obligations to those stakeholders early in the process of procuring an AI solution, managers can ensure that they are not caught in the knowledge gap, where many questions are asked of them without any answers. Instead, they can be more fully informed upfront about the benefits, risks, and mitigation strategies needed to use AI responsibly. Figure 4 illustrates the actual demand of stakeholders due to the moral and strategic obligations of firms. Current scholarship assumes firms are at point Z2, where the demands for knowledge from adopting firms center on minimal legal obligations and user acceptance, and the supply of AI knowledge through explainability and interpretability are similarly low. This article has served to more clearly state the existing knowledge demands based on obligations firms have to a broader group of stakeholders, moving firms up the y-axis to Z3. To meet these obligations, firms developing AI would then need to create a supply of knowledge to meet that demand and return the firm to equilibrium on point Z4.

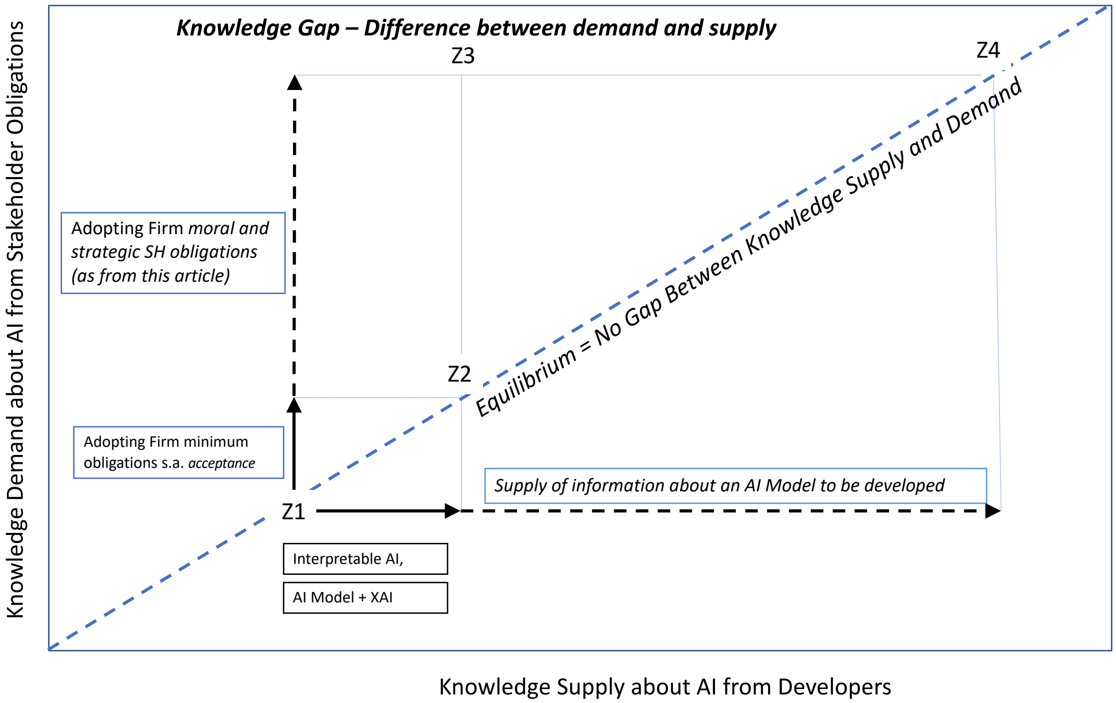

Knowledge Gap – Difference between demand and supply
Z3
Z4
Equilibrium = No Gap Between Knowledge Supply and Demand
Adopting Firm moral and strategic SH obligations (as from this article)
Z2
Adopting Firm minimum obligations s.a. acceptance
Supply of information about an AI Model to be developed
Z1
Interpretable AI,
AI Model + XAI
Knowledge Demand about AI from Stakeholder Obligations
Knowledge Supply about AI from Developers

Figure 4: Closing the AI Knowledge Gap.

IMPLICATIONS

In this article, we argue that firms have existing obligations to explain their actions to stakeholders, and the adoption of any technology cannot preclude a firm or manager from meeting those obligations. This argument has two implications: (a) how stakeholder obligations should inform not only the type of knowledge required but also design decisions and (b) whether or how 'black box' algorithms could ever be ethical to adopt within an organization.

New Design Criteria

The stakeholder obligations we identify here that drive a demand for knowledge should also guide how to make design decisions (Bleher and Braun 2023; Felzmann et al. 2020). AI developers make value-laden decisions in design – such as what training data to use, how to label data, which assumptions the model makes, defining outcome variables, and performance metrics, etc – that would be impacted by these same obligations (Martin 2022a). And these stakeholder obligations would guide the type of knowledge required by the firm about the ADM to ensure these obligations are met. For example, the obligation that the descision aligns with the goals of the firm should dictate the types of information to justify the manager is meeting this criteria, as argue above, but also the substnace of the design criteria. In other words, a manager must be able to justify decisions to meet regulations but also ensure that employment decisions meet the substance of those regulations. This paper's argument only addresses the type of knowelge required to prove to stakeholders that obligations are met. But the obligations can and should provide substantive guidance or procurement requirements driving design and development decisions for the elements of an intelligent AI model/agent: such as training data, model assumptions, outcome variables, input data, performance metrics, and even how the ADM is eventually used in practice (Asatiani et al. 2020). Figure 5 illustrates the many value-laden design decisions impacted by a single obligation.

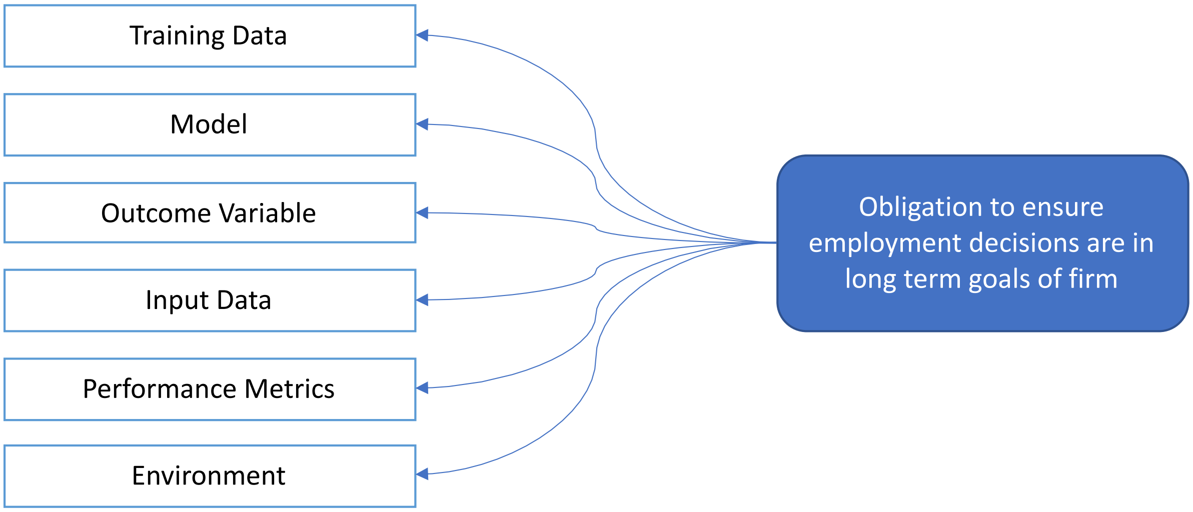


Figure 5: Value-laden decisions within the design of AI elements.

For example, the obligation to demonstrate that hiring does not discriminate along protected classes would require interrogating the choice of training data, the implications for the model chosen, the choice of the outcome variable, as well as the input data when in use, which performance metrics are used, and how the implmentation environment impacts the firm's ability to meet the regulatory obligations. In the Amazon case, the choice of the training data, (their own hiring practices over the past 10 years), proved to be a driver of the disparate treatment of applicants based on gender. In addition, because the program indexed on gender as the primary driver as to whether someone was hired or not, the program also began to hire people not even competent for the position because competence was not a factor in the training data to differentiate those hired or not hired.

Ethics of Black Box AI

Thus far, we have argued that stakeholders' demand for knowledge about the workings of AI models is greater than currently acknowledged within scholarship and practice. Further, firms will need to create a greater supply of knowledge to better meet the specific obligations of a stakeholder before adopting the AI system. An important implication of this analysis is that firms should not adopt AI models that do not provide the requisite knowledge to meet their stakeholder obligations. This argument runs counter to the inclination that adopting an AI program that is designed as a 'black box' could be a good business decision: where firms adopt AI first, assume the program is ‘accurate’ or ‘efficient,’ and not worry as much about justifying their decision to stakeholders.

The desire to justify the use of ‘black box’ AI is strong,[6] and it is worth considering how one *could* justify not understanding an algorithmic decision. First, there could be a class of possible inconsequential decisions where there might be no social justification required, no stakeholder obligations to justify firm decisions and little knowledge about why a decision was made is required. After going through the exercise above, no laws apply, no stakeholders are involved. For example, if a green background for an ad campaign was chosen over a blue background, a choice of pens for the supply cabinet, what type of pizza to order for lunch. There exist few situations within an organization where recommendations are typically adopted without any rationale provided. However, even such seemingly inconsequential decisions can be made for bad reasons. For example, if the color decision was manipulative or discriminatory or if the type of pen chosen only benefited a small group of people or harmed those in a protected class (Martin 2022b; Susser et al. 2019). In addition, as scale increases and more stakeholders are impacted, the firm may encounter different stakeholder values that require justification. For example, requiring green backgrounds might become an issue to discuss and explain in cultures where the color green takes on different meanings, such as in China or in Islamic countries.

One example of using AI with minimal explanation would be using image recognition to identify when a signature is needed for a government document (Asatiani et al. 2020). The case is illustrative, where the technical explanation may not be needed to meet any existing obligations. However, whether or not the program is effective in identifying the legally mandated signature would still be needed – in fact, the stakeholder obligation was the driving force for the creation of the image recognition program. Even within these seemingly inconsequential decisions, some degree of knowledge about AI is required to ensure the decision is in line with or does not conflict with the organization’s mission, values, and norms and with relevant stakeholder expectations.

[6] Google claimed that their large language model (Bard) spontaneously learned Bengali even with very few prompts in Bengali. “A Google AI model developed a skill it wasn't expected to have.” Google CEO Sundar Pichai said the company’s experts call this aspect of AI a “black box” – in response to the claim that Google’s large language model spontaneously learned Bengali. While Google did not provide the break down of the training data to undersan how Bard could learn Bengali, computer scientist Meg Mitchell did provide that breakdown and showed that the “black box” model was trained on the language Google claimed it spontaneously learned. https://twitter.com/mmitchell_ai/status/1648029417497853953

Second, and perhaps more interesting for the current approach to AI, would be a trusted authority scenario where a recommendation is made from a position of authority, and the rationale given to stakeholders is ***because an authority said so***. One can think of non-corporate situations where someone is in a position of authority, and their recommendations are taken without an immediate rationale, e.g., a head surgeon who has agency, organizational position, and expertise. A military commander may ask soldiers to do things they do not understand or agree with. A legal or financial expert may have more knowledge and suggest courses of action that someone does not understand.

Importantly, the actor in charge carries the moral force of laws and norms within a profession and a position of legitimacy. These experts are accountable to other institutions and people and are chosen to be in a position to enact the norms of society and the organizations they work within. A surgeon can be sanctioned and questioned by a board of surgeons, and a military commander can be asked to provide an explanation by their superior officers.[7] In these situations where the recommendation or direction is taken based on the authority and knowledge of the recommender, the authority figure is also responsible for the outcomes.

If AI is designed in ways that make it inscrutable, the firm that *develops* the AI program would need to take responsibility for how the program performs, who is negatively impacted by it, and any value that is lost by using the AI program. As of now, no firm developing AI has taken on such a financial, legal, and moral responsibility for their customers' use of the ADM. Instead, these developing firms want to be treated as if they are an authority but not take on the associated responsibility of being an authority.

The AI program designed as a black box carries the weight of a head surgeon in an OR but without the moral force of laws and norms, the position of legitimacy, expertise in the decision context, or, most importantly, without taking on the responsibility of the outcome of the recommendation. Recommendations taken and adopted without justification only work when the recommender (supervisor, officer, surgeon, etc.) also takes responsibility for the outcome.

Decisions made within a firm carry obligations to provide not only the recommendation but also a justification to stakeholders as to the fit with the firm's values, norms, and mission. These obligations do not disappear when an AI model is used to augment the decision. When

[7] In addition, defenses of 'obeying' orders are not the shield in court that they used to be. A historical survey of when and how the 'obedience to authority' defense has been used in the military. Solis GD. Obedience of orders and the law of war: judicial application in American forums. Am. U. Int'l L. Rev.. 1999;15:481.

using an AI program to read resumes, interpret interviews, and recommend who to hire, the hiring manager still has an obligation to provide a rationale as to why an individual was or was not selected. The knowledge needed to fulfill these obligations would vary based on the organization and decision and would become the design requirements for the AI program.

CONCLUSION

In this paper, we have argued that as managers and firms adopt AI to increase the efficiency and scale of various organizational processes such as hiring, firing, and optimizing resource allocation, their corresponding obligations to stakeholders to justify the decisions remain. When AI is designed as opaque and inscrutable, the gap between the information demanded by stakeholders and the information supplied by AI grows. This knowledge gap places managers in a risky position, increasing the odds of frustrating stakeholders, damaging trust, and increasing litigation.

To help managers uphold their obligations to various stakeholders and firms in the AI procurement process, we develop a list of questions that managers and firms can use to gain the benefits of AI adoption and minimize the risks of the related knowledge gap. Each category of questions helps managers better live up to their moral obligations to employees, customers, suppliers, shareholders, and the larger community.

Proponents of AI's potential in organizations may respond to this view by arguing that there are times when it is perfectly acceptable to obey orders from an inscrutable model, and with increased data and computing power, AI is like an authority that human managers should defer to. However, we argue that firms adopting AI models within ADM systems do so for decisions where a rationale is required for internal and external stakeholders. In other words, deciding to hire or fire someone comes with a corresponding obligation to justify the rationale to stakeholders.

REFERENCES

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, *20*(3), 973–989.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82–115.

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2020). Challenges of explaining the behavior of black-box AI systems. *MIS Quarterly Executive*, *19*(4), 259–278.

Benbya, H., Davenport, T. H., & Pachidi, S. (2020). Artificial intelligence in organizations: Current state and future opportunities. *MIS Quarterly Executive*, *19*(4).

Blair, M. M., & Stout, L. A. (2001). Trust, trustworthiness, and the behavioral foundations of corporate law. *University of Pennsylvania Law Review*, 1735–1810.

Bleher, H., & Braun, M. (2023). Reflections on Putting AI Ethics into Practice: How Three AI Ethics Approaches Conceptualize Theory and Practice. *Science and Engineering Ethics*, *29*(3), 21.

Buhmann, A., & Fieseler, C. (2021). Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society*, *64*, 101475.

Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1).

Burt, A. (2019). The AI transparency paradox. *Harvard Business Review*, *13*.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, *26*(4), 2051–2068.

Dastin, J. (2018, October 18). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Accessed 25 May 2023

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, *59*(2), 56–62.

Diakopoulos, N. (2020). Accountability, Transparency, and Algorithms. *The Oxford Handbook of Ethics of AI*, *17*(4), 197.

Dor, L. M. B., & Coglianese, C. (2021). Procurement as AI Governance. *IEEE Transactions on Technology and Society*.

Edwards, L., & Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, *16*, 18.

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, *26*(6), 3333–3361.

Freeman, R. E. (1984). Stakeholder management: framework and philosophy. *Pitman, Mansfield, MA*.

Freeman, R. E., & Phillips, R. A. (2002). Stakeholder theory: A libertarian defense. *Business ethics quarterly*, *12*(3), 331–349.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence-vs. integrity-based trust violation. *Organizational behavior and human decision processes*, *99*(1), 49–65.

Kim, T. W., & Routledge, B. R. (2020). Why a Right to an Explanation of Algorithmic Decision-Making Should Exist: A Trust-Based Approach. *Business Ethics Quarterly*, *32*(1), 75–102. https://doi.org/doi:10.1017/beq.2021.3

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable Algorithms. *University of Pennsylvania Law Review*, *165*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2765268

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., et al. (2021). What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473.

Lipton, Z. C. (2016). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *ACM Queue*, *16*(3), 31–57.

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57.

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

Martin, K. (2022a). Algorithmic Bias and Corporate Responsibility: How companies hide behind the false veil of the technological imperative. In K. Martin (Ed.), *Ethics of Data and Analytics*. New York: Taylor & Francis.

Martin, K. (2022b). Manipulation, Choice, and Privacy. *North Carolina Journal of Law & Technology*, *23*(3), 452–524. https://scholarship.law.unc.edu/ncjolt/vol23/iss3/2/

Martin, K. (2023). Predatory predictions and the ethics of predictive analytics. *Journal of the Association for Information Science and Technology*, *74*(5), 531–545.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., et al. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220–229). Presented at the Proceedings of the conference on fairness, accountability, and transparency.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI (pp. 279–288). Presented at the Proceedings of the conference on fairness, accountability, and transparency.

Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: nature, scope, and limitations. *Science and Engineering Ethics*, *27*(4), 1–30.

Nair, I., & Bulleit, W. M. (2018). Framing engineering ethics education with pragmatism and care: A proposal. In *2018 ASEE Annual Conference & Exposition*.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, *29*(3), 441–459.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Cambridge, MA: Harvard University Press.

Rader, E., Cotter, K., & Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency (pp. 1–13). Presented at the Proceedings of the 2018 CHI conference on human factors in computing systems.

Rucker, P., Miller, M., & Armstrong, D. (2023a, March 25). How Cigna Saves Millions by Having Its Doctors Reject Claims Without Reading Them. *ProPublica*. https://www.propublica.org/article/cigna-pxdx-medical-health-insurance-rejection-claims. Accessed 25 May 2023

Rucker, P., Miller, M., & Armstrong, D. (2023b, May 16). Congressional Committee, Regulators Question Cigna System That Lets Its Doctors Deny Claims Without Reading Patient Files. *ProPublica*. https://www.propublica.org/article/cigna-health-insurance-denials-pxdx-congress-investigation. Accessed 26 May 2023

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham L. Rev.*, *87*, 1085.

Someh, I., Wixom, B. H., Beath, C. M., & Zutavern, A. (2022). Building an Artificial Intelligence Explanation Capability. *MIS Quarterly Executive*, *21*(2), 5.

Speith, T. (2022). A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods (pp. 2239–2250). Presented at the 2022 ACM Conference on Fairness, Accountability, and Transparency.

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. *Internet Policy Review*, *8*(2).

Tigard, D. W. (2021). Technological answerability and the severance problem: staying connected by demanding answers. *Science and Engineering Ethics*, *27*(5), 59.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2021). The ethics of algorithms: key problems and solutions. *AI & SOCIETY*, 1–16.