JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 04/05 pp345−362
Volume 3, Number 4, December 2022
DOI: 10.23919/JSC.2022.0017

Fiduciary Responsibility: Facilitating Public Trust in Automated Decision Making

Shannon B. Harper and Eric S. Weber*

Abstract: Automated decision-making systems are being increasingly deployed and affect the public in a multitude of positive and negative ways. Governmental and private institutions use these systems to process information according to certain human-devised rules in order to address social problems or organizational challenges. Both research and real-world experience indicate that the public lacks trust in automated decision-making systems and the institutions that deploy them. The recreancy theorem argues that the public is more likely to trust and support decisions made or influenced by automated decision-making systems if the institutions that administer them meet their fiduciary responsibility. However, often the public is never informed of how these systems operate and resultant institutional decisions are made. A "black box" effect of automated decision-making systems reduces the public's perceptions of integrity and trustworthiness. Consequently, the institutions administering these systems are less able to assess whether the decisions are just. The result is that the public loses the capacity to identify, challenge, and rectify unfairness or the costs associated with the loss of public goods or benefits. The current position paper defines and explains the role of fiduciary responsibility within an automated decision-making system. We formulate an automated decision-making system as a data science lifecycle (DSL) and examine the implications of fiduciary responsibility within the context of the DSL. Fiduciary responsibility within DSLs provides a methodology for addressing the public's lack of trust in automated decision-making systems and the institutions that employ them to make decisions affecting the public. We posit that fiduciary responsibility manifests in several contexts of a DSL, each of which requires its own mitigation of sources of mistrust. To instantiate fiduciary responsibility, a Los Angeles Police Department (LAPD) predictive policing case study is examined. We examine the development and deployment by the LAPD of predictive policing technology and identify several ways in which the LAPD failed to meet its fiduciary responsibility.

Key words: trust; artificial intelligence; automated decision-making; recreancy theorem; fiduciary responsibility

1 Introduction

Automated decision-making systems are being rapidly deployed in the United States and internationally and affect the public in a multitude of positive and negative ways. Private and governmental institutions (i.e., societal institutions) use these systems to process information according to certain human-devised rules in order to address social problems or organizational challenges. These systems are often created using mathematical formulas or algorithms that are processed through computers to find commonalities among large datasets. For example, police departments have designed (with the assistance of data scientists)

• Shannon B. Harper is with the Department of Sociology and Criminal Justice, Iowa State University, Ames, IA 50011, USA. E-mail: sharper@iastate.edu.
• Eric S. Weber is with the Department of Mathematics, Iowa State University, Ames, IA 50011, USA. E-mail: esweber@iastate.edu.
* To whom correspondence should be addressed.
Manuscript received: 2022-07-10; revised: 2022-12-03; accepted: 2022-12-28

predictive policing algorithms to analyze massive amounts of pre-existing crime data to identify communities that have a high risk of crime, or past arrests or victimization data to identify individuals/groups who are likely to commit a crime or become a victim.

Some research suggests that the public lacks trust in automated decision-making systems and the institutions that deploy them[1, 2]. The recreancy theorem[3] argues that individuals are more likely to trust and support decisions influenced by automated decision-making systems if the institutions that administer them behave with integrity (i.e., fiduciary responsibility) and competency. However, often, the public is never informed of how these systems operate and resultant institutional decisions are made. A "black box" effect reduces the public's perceptions of automated decision systems' integrity and trustworthiness. Consequently, the institutions administering these systems are less able to assess whether the decisions suggested are just; and the public loses the capacity to identify and challenge unfairness, or the costs associated with the loss of public goods or benefits.

The current position paper examines fiduciary responsibility[1, 3] within the context of a data science lifecycle (DSL). There are many DSLs that affect individuals and the public at large, thus requiring institutional fiduciary responsibility. Examples of these DSLs include predictive policing[4−6], application processing (e.g., loans, school admissions, etc.), autonomous vehicles[7] and robotics[8], and government network surveillance and national security[1]. DSLs provide a holistic framework for describing processes and attributes of automated decision-making systems. A DSL has three layers: (1) a pre-processing layer, (2) a model building layer, and (3) a post-processing layer (see Section 2.1). Drawing from the recreancy theorem in quantifying the public's trust in automated decision-making systems, the current paper focuses on fiduciary responsibility within the third layer of the DSL. There is already a significant body of work to substantiate fiduciary responsibility within the early layers of DSLs (see Section 2.3). Our contribution is two-fold: (1) to analyze the notion of fiduciary responsibility within the third layer of a DSL, and (2) to assert that reducing the black box effect in that layer is necessary for institutions to meet their fiduciary responsibility (see Section 3). We discuss the role of fiduciary responsibility within DSLs, which provides a methodology for addressing the public's lack of trust in automated systems and the institutions that employ them to make decisions impacting the public (see Section 3.3). We posit that fiduciary responsibility appears in several contexts of a DSL, each of which requires its own mitigation of sources of mistrust. To instantiate our view of fiduciary responsibility within a DSL, a Los Angeles Police Department (LAPD) predictive policing case study is examined (see Section 4). We examine the development and deployment of predictive policing technology by the LAPD, and identify several ways in which the LAPD failed to meet its fiduciary responsibility. We further discuss actions and mechanisms which the LAPD could have utilized to meet its fiduciary responsibility.

The current position paper is situated in the relevant sociological literature concerning public trust in technological innovations. It provides a novel and potentially impactful framework to address and facilitate fairness, accountability, and transparency in automated decision-making systems, which spans the DSL workflow. Our analysis has a specific focus on building trust in the post-processing layer/stages. We also build on prior work to demonstrate how bias can manifest in the data acquisition, model building, and post-processing DSL layers/stages, requiring distinct mitigation strategies.

Important terminology

We will be using several phrases to describe, more or less, the same phenomenon that affects the public. These phrases are: (1) automated decision-making systems, (2) artificial intelligence (AI), and (3) data science lifecycles. We will be using them interchangeably, dependent upon context, though we acknowledge that they are not identical. "Automated decision-making systems" is a common phrase used in sociology literature[9−13] (but not exclusively[14]) and refers to institutional implementation of a mechanism—often without a human-in-the-loop—for making a decision and subsequently deploying an action that has an appreciable effect upon an individual or community. From our view, this is the best description of the systems we consider here when

viewed from the perspective of the public or stakeholders. As we will describe in Section 3, when the DSL occurs within a black box from the public's perspective, it is acting as an automated decision-making system. Artificial intelligence, for our purposes, is a methodology that (in part) mechanizes the automated decision-making system or appears as a stage within a DSL. As such, the AI moniker is narrower in scope than the overall pipeline that we have in mind for examining fiduciary responsibility. However, as it is very common in the literature, we still use it, particularly when we are referring to the work of other researchers. As emphasized in Ref. [15], AI also often operates within a black box from the perspective of the public. Finally, we use the phrase "data science lifecycle". "Data science" refers to methods and algorithms that interact with data, typically through acquisition, management, analysis, modeling, and reasoning[16−18]. As such, it encompasses more than statistics or data mining[19]. The term "lifecycle", or "pipeline", is becoming more common in the literature[16, 20−26]. We describe our usage of the phrase in Section 3. To emphasize the point here, DSL is meant to be an encompassing term that includes both AI and automated decision-making systems. Ultimately, using the notion of fiduciary responsibility, we will demonstrate that to facilitate public trust in these systems, much of the DSL should operate in view of stakeholders.

2 Conceptual Framework: Fiduciary Responsibility

2.1 Recreancy theorem and fiduciary responsibility

As conceptualized by Sapp et al.[3], the recreancy theorem argues that the public's trust in public and private societal institutions is explained by their perceptions of the institution's competence (i.e., skill, ability, and experience), and their confidence that the institution will behave with integrity (i.e., honesty and ethical standards), also known as fiduciary responsibility[27]. Benevolence is a third central component of public trust, which involves the perceived extent to which the institution is concerned about citizens' welfare. Multiple scholars have argued that when societal institutions (with a wide spectrum of roles and responsibilities) fail to build trust among the populations they serve (i.e., reflect recreancy), society's ability to function is detrimentally effected[1, 27−29]. Additionally, some have argued that trustworthiness rather than trust signifies public opinions of societal institutions' behaviors[1, 30, 31].

Sapp et al.[1] (see also Ref. [27]) defined trustworthiness as institutional behaviors that give citizens reason to have confidence in their performance. Interpersonal trust is embedded in the recreancy theorem where there is a perceived connection between a societal institution and individuals in the population it serves[1]. When the public perceives an institution as meeting their expectations of competent, honorable, and benevolent performance, interpersonal trust between the two is established[32−35], and recent research affirms this contention[1, 31, 36, 37]. Such perceptions influence whether the public trusts and supports technological innovations (such as automated decision-making systems) proposed or administered by these institutions[32, 33, 38−40]. Additionally, recent research reveals that the public is more likely to trust social institutions when their newly administered AI does not negatively affect social justice—i.e., protects the interests of vulnerable/marginalized populations[1, 37, 41, 42].

The degree to which the public trusts institutions in their administration of automated decision-making systems often influences whether those systems are abandoned or used long-term[1, 3]. Fiduciary responsibility[1] is an integral tenet of the recreancy theorem and refers to public perceptions of the integrity demonstrated by societal institutions, which influences individuals' trust and support for automated decision-making systems administered by those institutions. In the current paper, we specifically examine fiduciary responsibility in societal institutions' development and administration of automated decision-making systems because embedded processes of data collection, data modeling, and prediction output influence whether the public will perceive those institutions as having integrity. Following automated predictions, some action typically occurs. For example, in a system that processes loan applications, the system decides whether to approve a loan application after making a prediction on whether an applicant will likely repay the loan. The prediction and subsequent decision are based on a data collection method and a model developed by the designer of the system. Such automated decision-making systems are

part of a larger group of data driven processes that are often called DSLs or sometimes, "data science pipelines". Societal institutions that administer algorithm-driven automated decision-making systems with honor—e.g., concern for and attention toward minimizing bias/racism and privacy infringement—are more likely to be trusted by the public[1]. We will show that automated decision-making systems that operate within a "black box" (where data scientists and institutional staff make system development and administration decisions "in the dark") absent public technological knowledge, informational awareness, and scrutiny, hinder an institution's efforts to meet fiduciary responsibility, and consequently establish trust.

2.2 Fiduciary responsibility and public trust in technical innovations

The recreancy theorem, as applied specifically to technical innovations, delineates three dimensions of public acceptance of technical innovations. As stated in Ref. [1], the recreancy theorem: "complements technology adoption models in that it focuses upon public assessments of innovations as they are managed by societal institutions, thereby providing conceptual congruity between technology adoption and public assessments of institutional competency and integrity."

To that end, Refs. [3, 43] argued that the public is more likely to trust and accept the implementation and use of technologies that pose a risk to their welfare or interests provided they meet the following requirements. First, they (the technologies and/or institutions deploying those technologies) must possess **technical competency**, meaning that they must have the capability to perform the analysis and/or actions required for the intended purpose of the technologies. Second, they must have a **public benefit**, meaning that these technologies should improve upon a specific problem of public concern upon deployment. And third, they must meet **fiduciary responsibility**, meaning that the algorithms are designed and employed with integrity, and are in fact performing in the way that they are intended, without disparate impacts or misuse by agents.

Each of these dimensions warrants investigation with respect to the public's trust in AI deployment in particular. We describe in Section 2.3 prior works that have related themes, especially methods for formalizing, analyzing, and quantifying technical competency. As such, we will not address that dimension here, nor will we consider the public benefit aspect of the recreancy theorem. We will focus on fiduciary responsibility, building upon recent research findings that fiduciary responsibility is of particular importance among the public, especially for the protection of vulnerable populations. Scholars suggest that institutional administration of technological innovations raises multiple concerns specific to integrity, including invasion of privacy, loss of personal health data, and unfair monitoring or targeting of particular individuals or groups, which may sometimes involve racial bias[1, 44−48].

Further, Ref. [1] discussed a conceptual understanding of public opinions of network surveillance and empirically documented public demand for network surveillance that fosters goals of social justice more than goals of self-interest. The findings are based on a nationwide survey of adults concerning governments' use of network surveillance. Additionally, the utilization of a technological innovation is often perceived as socially just when it protects the rights of marginalized/vulnerable populations such as people of color, women, and LGBTQ+ individuals[1].

The recreancy theorem asserts that the public's acceptance of technical innovations depends on the public's perception that institutions fulfill fiduciary responsibility while deploying those technologies. Therefore, it is incumbent upon institutions to ensure that the public is confident that fiduciary responsibility is met by the technologies in use. By utilizing the formalism of a DSL, we will argue that this can best be done through transparency at multiple stages of the lifecycle, especially those stages that occur in the third layer identified as interpretation and communication (see Figs. 1 and 2).

2.3 Related work

Public trust in automated decision-making systems or artificial intelligence is, *prima facie*, both important to establish and difficult to formalize. For example, Refs. [49−51] put forward competing formal definitions of trust, either interpersonal or institutional, and these definitions have various dimensions. Several high profile institutions[52, 53], just to name a few, acknowledge the importance of establishing trust in AI

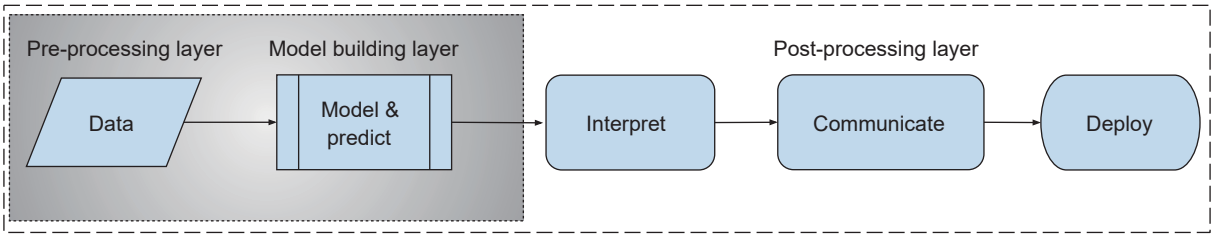

Fig. 1 A reduced model of a DSL. The pre-processing and model building layers operate within a black box from the perspective of the constituent/stakeholder, while the post-processing layer operates openly from the perspective of the constituent/stakeholder.

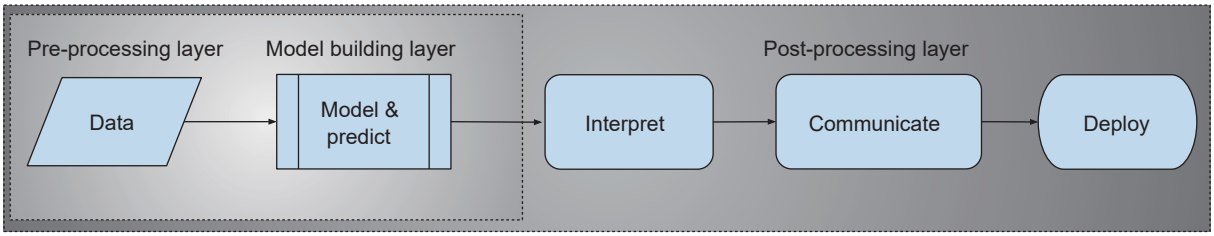

Fig. 2 A reduced model of a DSL. Here, all three layers operate within a black box from the perspective of the constituent/stakeholder, while the post-processing layer operates openly from the perspective of the agents utilizing the DSL.

while using the term in a colloquial sense. Early efforts to formalize trust in various forms of digital interactions appear in Refs. [54, 55].

Recent work has endeavored to describe trust in AI in several different ways on a technical basis. Several authors posited a formal definition of trust in AI by drawing on prior work in formalizing interpersonal and institutional trust[56, 57]. Other authors proposed methods for quantifying, or establishing, trust through established legal or public policy structures[15, 58]. Still, others distinguished between trust in AI versus trustworthy AI[59]. In the context of public trust in AI, there are multiple ways in which our use of the recreancy theorem in general, and fiduciary responsibility in particular, are related to these other works. We describe those relations here.

It is well understood that public trust in AI has multiple dimensions. The three facets of the recreancy theorem—competency, benefit, and fiduciary responsibility—are reflected in others' investigations into the question of how AI can be trustworthy. Indeed, Ref. [56] found similar dimensions as the recreancy theorem: "To investigate whether a global agreement on these questions is emerging, we mapped and analyzed the current corpus of principles and guidelines on ethical AI. Our results reveal a global convergence emerging around five ethical principles (transparency, justice and fairness, non-maleficence, responsibility, and privacy), with substantive divergence in relation to how these principles are interpreted."

Toreini et al.[57] described trust in AI as distinct from trustworthy ML by utilizing the ABI framework—Ability, Benevolence, and Integrity—posited by Ref. [60] to model organizational trust. We note that Mayer et al.'s ABI framework and the recreancy theorem run parallel to each other in their dimensions, while Toreini et al.'s notion of trustworthy as a part of trust reflects the technical competency of the recreancy theorem. This is borne out by several related notions such as trustworthy AI and human-AI trust.

In general, trustworthy AI refers to the competence of the AI algorithm, with competence defined in terms of a contract. This is precisely how Jacovi et al.[59] defined the notion: "an AI model is trustworthy to some contract if it is capable of maintaining this contract." Similarly, Ref. [61] described trust facilitated through a commitment (i.e., a contract). This commitment also involves expectations by the trustor about the competence and willingness of the trustee. Knowles and Richards[15] developed a foundation for trust in AI through a transparent and understandable regulatory system. They argued that a regulatory system can potentially bridge the gap between trustworthiness and trust. Such a regulatory system can be a mechanism for extending the public's trust from

acceptance that a particular AI meets the required technical competence to meeting the required fiduciary responsibility as well.

Jacovi et al.[59] further proposed a formal definition of human-AI trust via adapting the sociological definition of interpersonal trust. The definition has several dimensions, including trustworthiness of the algorithm and warranted trust possessed by the human who is at risk of the AI's actions. They posited that an algorithm is trustworthy if it can uphold a contract; and a human possesses warranted trust of AI if the human has reason to accurately anticipate the impact of the AI's decisions. The notion of trustworthy put forth in Ref. [59] need not require an individual to understand the inner-workings of the algorithm.

Others argue for facilitating public trust in AI through an understanding of the workings of the algorithms, either by individuals that are subject to AI or by experts who can vouch for the competency of the algorithms. A large body of literature discussed explainable AI (XAI) as one method for ensuring trust between users and AI[62, 63]. Documentation, such as AI factsheets[58, 64] (for example, the European requirements for factsheets[53]) or declarations of conformity[65], facilitate auditing of AI, and thus assist in building public trust. Both of these mechanisms are designed to meet the perception of the algorithms' technical competency among the public and consequently establish trust.

Our discussion here concerns (predominantly) the public's trust in technological innovations as stakeholders within the environment that the innovation is deployed. As such, the recreancy theorem is one of multiple technology adoption models (TAM). A TAM that centers on user acceptance was introduced by Davis[66, 67]. In our analysis, the users are the institutions deploying the technology, and thus the recreancy theorem is the relevant model for understanding public acceptance rather than user acceptance.

3 Public Trust in Data Science Lifecycles

We seek to formalize the notion of fiduciary responsibility within a DSL as a way to facilitate public trust in systems that utilize AI. We will adapt the formal description of a DSL as discussed in Ref. [16], which decomposed a DSL into three layers and each layer into multiple stages. We will argue that formalizing a common specification of a DSL can facilitate multiple mechanisms through which to establish public trust. Finally, we will show how the requirements of fiduciary responsibility are embedded within a DSL, and how the common specification of a DSL can facilitate the meeting of those requirements by organizations and institutions that design and deploy them.

3.1 DSL formalism

Data driven processes follow varied forms and take on many functions, but recent work has been made to accurately describe these processes in ways that facilitate our discussion of fiduciary responsibility. Reference [16] comprehensively described and classified DSLs. The study found that some DSLs have very few stages, as few as 4 or 5, while others have as many as 11 (as suggested by Ref. [26]). Many of the publicly available DSLs identified in the study were used only for academic or competitive (e.g., Kaggle) purposes and were not in fact deployed in a real-world environment. These DSLs typically lacked the stages that form our focus, i.e., those that occur after the prediction stage within the DSL.

Following the specification from Ref. [16], a DSL has three layers: a pre-processing layer, a model building layer, and a post-processing layer. The pre-processing layer consists of three stages: data acquisition, data preparation, and data storage. The model building layer consists of five stages: feature engineering, modeling, training, evaluation, and prediction. Finally, the post-processing layer consists of interpretation, communication, and deployment. The DSLs that we consider should explicitly have all of the 11 stages laid out in Refs. [16, 26] (see Section 3.3), but our focus will be on the latter stages within the overall process. We therefore depict the DSL (in a reduced form from Ref. [16]) for our purposes in Figs. 1 and 2.

The stages within the post-processing layer are described as follows in Ref. [16].

"**Interpret:** The prediction result might not be enough to make a decision. We often need... post-processing to translate predictions into knowledge.

Communicate: ...we might need to communicate with the involved parties (e.g., devices, persons, and systems) to share and accumulate information.

Deploy: The built DS solution is installed in its problem domain to serve the application..."

These descriptions of the stages within the post-processing layer, as well as those of the stages in the other layers, refer largely to DSLs that are fully automated. However, in the contexts that we are considering—predictive policing, application processing, etc.—there often is a human-in-the-loop (as described in the case study presented in Section 4.1). Specific to the LAPD case study we describe in Section 4, hot spots predictive policing uses an algorithm in the pre-processing layer that attempts to predict high crime areas. Next, police personnel interpret the prediction to deploy crime-reducing resources, e.g., an increased number of foot-patrols in that area. As a result, our view of the post-processing layer is distinct from, though still related to, the view in Ref. [16]. For our purposes, we describe the stages of the post-processing layer as follows.

Interpret: The prediction requires evaluation by an agent in order to make a decision or recommendation. The evaluation may involve the context in which the prediction occurs and/or additional information that is not available to the model building layer of the DSL. Here, the agent can be a cyber or physical system, but is likely to be human.

Communicate: The agent(s) that evaluate the prediction communicate the interpretation to other agents for the purposes of making a decision. These other agents can also be cyber, physical, or human.

Deploy: Based on the prediction and interpretation, a decision is made. Subsequently, an action is taken; usually this action is taken in the physical world and may involve the deployment of resources.

Returning to our predictive policing example, we note that the public perceives all layers and stages of the DSL occurring within a black box—this occurred specifically in the implementation by the LAPD. Most notably, the post-processing layer, which operates with humans-in-the-loop, is obscured from public input, scrutiny, or accountability.

3.2 Formalizing DSLs to facilitate public trust

There are multiple mechanisms for facilitating public trust in AI through formalizing DSLs. We will see shortly how following a formal DSL framework can promote the public's perception of institutions' fiduciary responsibility. Independent of our thesis regarding fiduciary responsibility, the DSL framework can assist in establishing public trust in the technologies that are deployed.

Establishing a common DSL format can facilitate consistent development and maintenance, as well as the production of useful DSL documentation when considered as a software engineering endeavor[16]. In turn, the common DSL format and concomitant documentation can facilitate trust in AI-as-an-institution, as argued in Ref. [15]. The DSL framework provides multiple specifications that can be utilized for developing regulatory infrastructure, contract formation, fact sheets, and other "structural assurances" to facilitate the trustworthiness of AI deployment.

In addition, a common DSL format can also reduce the black box effect, akin to the assertion by Ref. [68], through providing a description of the internal workings of the black box in order to rectify the knowledge gap. The common DSL format decomposes a larger black box into smaller black boxes (as depicted in Figs. 1 and 2), some of which can be open to the public, and others of which can be subject to expert auditing through declarations of conformity. While the public may not understand the black boxes, or even the DSL framework proposed in Ref. [16], the commonality can help it conceptualize the documentation requirements and accountability to which DSLs are subject.

We have mentioned just a few elementary ways in which a formal DSL framework can facilitate public trust in AI. We have not fully explored this consideration and there is much more work to do in this regard, but such work is outside the scope of the present paper. Scholars such as Sapp et al.[1] used structural equation modeling to quantify public trust, finding that considerations of technical efficacy and social justice are significantly and equally associated with public trust in and support for government-administered network surveillance (see also Ref. [43]).

3.3 Fiduciary responsibility within data science lifecycles

We argue that meeting fiduciary responsibility within a DSL requires the post-processing layer to operate in an open box capacity. The purpose of opening the post-processing layer is to provide stakeholders a context for understanding decisions made, which may build public

trust in the action(s) being deployed/implemented. An additional potential benefit of the open box operation is that the public has a means to hold institutions accountable for the decisions made by DSLs. We emphasize here that the open box operation is a necessary but not sufficient condition for meeting fiduciary responsibility.

As described in Ref. [16], some DSLs can be fully automated, such as bank loan application processing systems. In such a fully automated system, from the perspective of the person subject to the decision of that system, it operates entirely within a black box (as depicted in Fig. 2). Naturally, the person is likely not provided any information or context regarding the decision and therefore cannot understand the decision process. Absent other mechanisms—such as documentation or auditing by trusted experts—fiduciary responsibility cannot be met in this regime, which in turn precludes the establishment of trust by those persons subject to the DSL.

We consider here DSLs that are not fully automated, but involve humans-in-the-loop, particularly those DSLs in which humans appear in the post-processing layer. For example, we re-imagine the loan application processing DSL in which the prediction of the model is given to a (human) banking specialist. This person will interpret the prediction, perhaps utilizing additional information not available to the model or placing the prediction within a larger context. The specialist then informs (communicates) the applicant of the decision, at which time an action (deployment) is taken, e.g., the loan is fulfilled, or the application is closed.

We explore the multiple facets of fiduciary responsibility with the DSLs we have just described, focusing on the instantiation of fiduciary responsibility within the post-processing layer. Before doing so, however, we want to acknowledge that fiduciary responsibility manifests within all layers and stages of DSLs. Much work has been done already to substantiate fiduciary responsibility within the pre-processing and model building layers of DSLs as we described in Section 2.3, and yet more work remains.

For the sake of context, we mention several facets of fiduciary responsibility that appear within the early layers of DSLs, while emphasizing that our comments here are not exhaustive. At the pre-processing layer, individuals who are subject to a DSL expect that data associated to them are accurate and will be kept private. It is likely that informed consent is required at this stage. Individuals also expect that data associated to them and others collectively do not contain bias, or put them at higher risk for adverse decisions or actions. As part of the model building layer, individuals expect that the model accurately analyzes the data without introducing spurious effects or amplifying bias[9]. In addition, at the prediction stage, individuals expect to be able to anticipate the impact of the model's prediction[59] and subsequent decision. Moreover, the individual expects that the DSL is not being misused or abused by the institution or its agents.

Let us now turn to the several stages within the post-processing layer. Again, we are considering DSLs for which these stages are an integral part, since a decision and subsequent action are necessary for the public to be affected by the DSL, independent of whether they are aware of them. Our examples of DSLs above indeed incorporate these stages in some form or other.

The first stage of the post-processing layer is "Interpret", by which we mean an actor interprets the prediction made by the model. As mentioned previously, this actor could be cyber or human, though we focus on a human agent here. Fiduciary responsibility requires the agent to make an interpretation which is honest, ethical, and just. The interpretation should reflect the institution's values, mission, and goals, as well as uphold the rights and interests of stakeholders and those impacted by the institution's operations. The interpretation should be understandable to those who are subject to the decision of the DSL, and, as we shall discuss shortly, amenable to communication to individuals, stakeholders, auditors, etc. Interpretations should be well-documented and archived. As a reflection of these criteria, an interpretation serves dual purposes.

The initial purpose of the interpretation is to provide the decision-maker with a fuller understanding of the prediction. For example, from a data science perspective, the prediction may have an associated confidence level or indicate the most relevant features of the input data leading to the prediction. The interpreter can provide context to these additional pieces of information for the decision-maker: as Dobbe et al.[9] argued that "machine learning models should *facilitate rather than replace* the critical eye of the human expert" (emphasis in the

original).

The subsequent purpose of the interpretation is to provide other entities with a vested interest in the DSL[69]—such as the individual (or community) subject to the decisions of the DSL—with the context or the rationale for the decision. We note that the context for the decision-maker and that for the individual are assuredly different, (for example, see Ref. [69] which identified function roles that inform the nature of an interpretation or explanation) but likely have much overlap. This is because, in both cases, ultimately a human will want to have some understanding of the mode's prediction. An interpretation, while it need not explain the inner workings of the algorithms involved, does have the potential to assure individuals that the algorithms are operating with competency, and is even able to provide individuals a foundation for harmonizing the decision with their anticipation of that decision[59].

Proper anticipation of the DSL's output is a potential foundation for building trust between an individual (or community) and an institution deploying the DSL[59]. We have previously described how XAI is one commonly identified methodology for establishing this foundation. In our DSL, the interpretation can be informed by XAI when utilized within the model building layer; conversely, a well-formulated interpretation can counteract the lack of explainability when a model does not utilize XAI.

We further contend that the interpretation of the DSL output should be driven by values in addition to technical and explanatory considerations. Value-laden interpretations can address epistemological issues related to fairness, accountability, and transparency. As Dobbe et al.[9] further argued, questions of fairness "illuminate the range of places in the machine learning design process where issues of *epistemology* arise: they require *justification* and often *value judgment*" (emphasis in the original).

Interpretations, which are value-laden, that are embedded as a formal stage of the DSL also situate the causes and effects of the DSL within a broader context of "the human element", personalizing both the individual affected by and the institution deploying the DSL.

Our formalism of DSLs, with "Interpretation" firmly embedded within the overall process, contextualizes epistemological questions, such as "Interpretable to whom?" or "For what purpose?" as asserted by Kohli et al.[70] As we have already identified, the interpretation has several potential audiences: (1) anyone else downstream within the DSL, and (2) those affected by the DSL's output. These audiences then proscribe the purpose(s) of the interpretation: (1) to help actualize the ultimate goal of the DSL, and (2) to assure stakeholders that the DSL's output is properly anticipated.

The penultimate stage of the DSL is "Communicate", yet from the viewpoint of fiduciary responsibility, "Communicate" is the most crucial of all stages. Indeed, as the recreancy theorem measures the public's perception of the integrity, honesty, and justness of an institution and its use of DSLs, full and open communication by the institution is of paramount importance for meeting its fiduciary responsibility. Indeed, communicating the interpretation of the DSL's output to affected individuals or communities actualizes the secondary purpose of the interpretation, thereby establishing a potential foundation for trust. As observed in our case study on the LAPD's use of a predictive policing DSL in Section 4, a lack of communication can lead to the public not trusting in the institution's deployment of a DSL.

Communication opens the black box operation of the DSL at least to the extent that, if successful, individuals and/or the public at large can understand the rationale, if not the mechanism, for the decision and subsequent action. This opening of the black box is shown in Fig. 1, and is premised on the argument in Ref. [15] that "lack of public trust in AI has little to do with people's inability to understand how AIs work; rather it is a response to an awareness of a lack of structural assurances of the trustworthiness of the AIs pervading society".

Hence, we envision a DSL with the early layers still operating within a black box, but the post-processing layer operating openly. We argue that this regime of a DSL operation can greatly advance the institution's efforts to meet its fiduciary responsibility as well as provide the potential for establishing "structural assurances of trustworthiness" that the public will require for accepting the implementation of a DSL.

The content of the communication by the institution that deploys a DSL consists of multiple aspects. The

form of the communication is likely dependent upon the institution and its values, the nature of the decision itself, and the affected party (e.g., whether an individual or a community). We formulate the communication in part around the rhetorical "Five Ws" (who, what, when, where, and why). The institution informs the affected party of (What:) the decision made and the subsequent action that was/will be taken; (Why:) the interpretation of the prediction within the larger context of the DSL and the institution's mission, the rationale for the decision based on the model's prediction, and the justification for the action based on the prediction and the decision; (How:) what data were used to make the prediction, how the data were utilized, how the input data as well as the model output were interpreted, and the values that informed the interpretation of the model output.

The final stage of the DSL, which we have referred to as "Deploy", is the point at which an individual (or a community) is ultimately affected by the DSL. This is not to say that an individual is not ever affected in prior stages—in fact, this is a distinct possibility, e.g., lost privacy—but this stage manifests the ostensible *raison d'etre* of the DSL. We use "Deploy" to be consistent with Ref. [16], though our usage is distinct. Here, we think of an institution choosing whether and how to deploy resources—for example, manpower or finances; but at a more basic level, this stage refers to the institution implementing an action.

Fiduciary responsibility requires an institution and its agents to employ honest, ethical, and just actions. In the context of automated decision making, the actions need to be well-founded in the model, prediction, and interpretation, meaning that the actions are justified through accurate models, correct predictions, and valid interpretations. The DSL formalism provides a framework for ensuring that fiduciary responsibility is met at each stage of the lifecycle, particularly through documentation and auditing[15]. In addition, the DSL framework instantiates meeting fiduciary responsibility through both technical[6, 59] and ethical[1, 56] dimensions. We note that some of the issues associated to technical and ethical concerns are context dependent.

An institution is required to employ actions that protect the rights of the affected individuals, particularly those of vulnerable populations. In the course of doing so, institutions likely need to document the history of actions that have been taken, and ensure that the actions are just in specific and in aggregate. Auditing of the actions must occur by trusted experts. The DSL framework provides the institution with a systematic (i.e., system-level) method for identifying issues of fiduciary responsibility and documenting the methods for addressing the issues, both during and after implementation of the DSL. Institutions can use several methods for this documentation as developed by others. For example, institutions can publish for public consumption declarations of conformity (DoC)[65] or factsheets[53], giving stakeholders the opportunity to evaluate an institution's fiduciary responsibility. Likewise, institutions can utilize contracts[59] for the benefit of stakeholders as well. This particular aspect of the theory requires additional work beyond the scope of this article.

4 Predictive Policing DSLs: Benefits, Risks, and Public Trust

Over the past 14+ years, multiple urban police departments across the United States have sought and utilized algorithm-driven predictive policing technologies that evaluate massive volumes of historical crime/arrest data to predict high crime geographies/places or crime prone individuals, which help police leadership decide where and how to deploy officer resources. Predictive policing DSLs were pioneered by the Los Angeles Police Department (LAPD) in the 2000s and quickly spread to several major cities across the US[71]. There are two forms of predictive policing technology: place-based and person-based[71]. Place-based predictive policing is the most extensively used method and leverages pre-existing crime data to identify places and times with a high probability of crime. Alternatively, person-based predictive policing looks for risk variables like previous arrests or victimization trends to identify individuals or groups who are likely to commit a crime or be a victim of one.

Proponents of predictive policing technology assert that benefits of the technology include assisting police in forecasting crimes more objectively, precisely, and effectively than traditional policing methods and investigation techniques[4, 71]. Predictive policing is intended as an automated tool to reduce primary reliance on officer instincts to forecast crime[71], thus

increasing officer safety and the accuracy of crime prediction. Technology designers claim that they are not only capable of substantially reducing violent crimes such as murder, aggravated assault, and robbery, but also of removing bias from police decision-making[71, 72]. However, claims that predictive policing reduces crime have been disputed[4, 73−75].

The majority of risks associated with predictive policing technologies are related to the black box effect described in Section 3. Predictive policing DSLs rely on previous crime data that are often incomplete due to a large percentage of crime being unknown or unreported, and/or racially biased due to the disparate arrests of African American and Hispanic people when compared to Whites across time[76, 77]. Some scholars argue that racism and bias are systemically entrenched in the criminal justice system (CJS), facilitating the disproportionate mass incarceration of people of color, and influence police practices, policies, and behaviors on the ground[78−80]. Black, Brown, and low-income communities have been over-policed historically due in part to the social acceptance of racism and high crime rates that are often related to poor structural conditions and a lack of access to resources associated with systemic inequities in the CJS and society at large[78]. Considering these arguments, the methodology through which officers collect the data impacting the pre-processing stage of the DSL may be influenced by systemic racism and racial bias in policing; and predictive policing algorithms (model building layer) rely on such data to generate place or person based predictions in the post-processing stages of the DSL[79], which can perpetuate or reinforce historical prejudices in policing practices and policies[71, 79, 81−84]. Compounding the severity of such concerns, the manner in which police develop and administer predictive policing DSLs often lacks mechanisms to hold departments accountable for the interpretation of predictions and the decisions made/actions taken based on those predictions[71, 79, 81−83, 85, 86]. In the same vein, algorithmic predictions (DSL post-processing layer) can influence how police officers view the neighborhoods they are patrolling, and the ways in which they perceive individuals' criminal propensity within those (primarily Black and Latino) communities[87, 88], which ultimately may affect probable cause for arrest decision-making.

Some scholars and activists also posited that predictive policing DSLs facilitate increased and unjustified police stop, search, and seizure decision-making that can violate the Fourth Amendment of the US constitution[71, 87]. Privacy concerns of predictive policing DSLs include eroding public anonymity through expanding webs of surveillance in the US, and creating networks of personal information that can be shared across police departments, accessed through illegal computer hacking or system breaches, or mishandled by officers[89]. Arguably, these outcomes negate any potential benefits of predictive policing technology[79]. Such risks have resulted in public (as well as data scientist) protests, boycotting, and privacy protection activism across the US where predictive policing technology has been proposed or used[85, 90−93].

4.1 Black box effect: A case study of the Los Angeles Police Department (LAPD)

The LAPD began testing and implementing a person-based predictive policing DSL known as "Operation Laser" (referred to as Los Angeles Strategic Extraction and Restoration (LASER)) in 2011. LASER identified repeat offenders, and produced bulletins with their photos and physical descriptions so law enforcement could find/identify those individuals to prevent their future criminal activity[86]. Examining LASER's operation within the DSL framework, the LASER algorithm utilized criminal history data to identify individuals most likely to commit a violent crime as part of the pre-processing layer of the DSL. "Chronic offenders" were ranked using a point system where factors like gang membership, number of perpetrated violent crimes, and interactions with officers were algorithmically assessed within the model building and prediction stage (second DSL layer) of the DSL. Subsequently, those with higher numbers of "points" were placed on chronic offender bulletins within the "Interpret" stage of DSL post-processing that were distributed to officers during the DSL post-processing "Communicate" stage. These bulletins provided law enforcement with the identifying information necessary to specifically target (i.e., approach) those on the list as part of the "Deploy" stage of the DSL post-processing layer.

Andrew Ferguson, a law professor and nationally renowned predictive policing expert, explained to CBS News[94] that "the LASER program was designed on

the metaphor that they [the LAPD] were going to, like laser surgery, remove the tumors, the bad actors from the community…That idea, offensive as it is, was an idea of using some kinds of algorithms to identify risk". Implementation of LASER led to public outcry specific to DSL predictions being used as a legal veneer for police brutality, mistreatment, and racial profiling against people of color[81, 95, 96]. The LAPD Inspector General conducted an internal audit of LASER in 2019 (eight years post-implementation), and found that Latinos/as and African Americans made up 84% of "active" chronic offenders. The audit revealed numerous inconsistencies (relevant to the model building and "Interpret" stages of the DSL) where the LASER algorithm identified and labeled individuals as "chronic offenders"[94, 97]. More specifically, 44% of labeled "chronic" offenders had never been arrested or only had one arrest for some type of violent crime, and nearly 10% had no "quality interactions" with law enforcement[94, 97]. The program was discontinued in 2019 following the audit.

The LAPD also contracted with PredPol in 2011, which used historical property crime data to produce "hot spot" predictions with a high likelihood of vehicle theft and burglary[71]. PredPol applies an "earthquake" crime prediction method which—like earthquakes and aftershocks—smaller crimes lead to bigger crimes and occur in near proximity to one another[85, 98, 99]. Like LASER, PredPol was in part intended to prevent subjective judgments and implicit bias as part of officer deployment decisions[100]. However, activists and anti-predictive policing community members maintain that the overwhelming bias of PredPol—i.e., hot spot map predictions identifying primarily Black and Brown neighborhoods—renders it unreliable and corrupt, and thus cannot be trusted and must be entirely dismantled[101−103]. In relation to the post-processing layer of the DSL, when police leadership interprets PredPol's problematic hot spot predictions and thereafter decides to assign higher or increasing numbers of officers ("Communicate" stage of the DSL) to patrol African American and Hispanic/Latino/a/x communities ("Deploy" stage of the DSL), the likelihood of civil rights and civil liberties violations increases. According to the recreancy theorem[1], such actions within a black box do not reflect fiduciary responsibility, which further helps to explain Los Angeles residents' criticisms of, and wariness about, the department and its use of predictive policing. We envision an idealized DSL in which a (human) policing/data specialist receives the output of the model and interprets this output within a broader context. This interpretation is communicated to someone with the authority to implement policing strategies who decides how to utilize the information and then determines the deployment of resources.

Anti-predictive policing protests, advocacy organization mobilization, and academic criticism began to escalate in 2016, and 17 groups, including the American Civil Liberties Union (ACLU) and the National Association for the Advancement of Colored People (NAACP), signed a widely circulated statement indicating their concerns about the reinforcement of racial bias associated with predictive policing technologies[81] and the lack of transparency about development and use of the DSLs from the institutions that administer them[71]. PredPol was discontinued by the LAPD in 2020; the department claimed that this action was taken because of COVID-19 financial constraints[71, 104].

4.2 Manifesting public mistrust in predictive policing

Referring back to Section 3, depending on the nature of the predictive policing prediction output, police personnel interpret (i.e., evaluate) the prediction, communicate the evaluation with other personnel, and ultimately deploy (i.e., take action) police resources accordingly. Fiduciary responsibility manifested in the DSL framework does much to explain public mistrust of predictive policing as implemented by the LAPD. Public mistrust was largely associated with the LAPD's lack of transparency as the department developed and administered the DSLs across the pre-processing, data modeling, and post-processing layers of the DSL (see Ref. [71]). The LAPD began utilizing LASER and PredPol as mechanisms of crime control absent disclosures to the public about the basics, intricacies, development and administration, or decision-making associated with the technologies[86], which can be explained by the black box effect depicted in Fig. 2 rather than in Fig. 1.※

More specifically, the minimal information that the

※It is important to note that this argument is formulated based on the authors' search of media and government press release databases.

LAPD did share with the public lacked thorough and transparent details on the multiple layers across the DSLs, such as the types of data utilized (i.e., pre-processing layer), how the algorithm(s)/technologies were developed and produced predictions (i.e., model building layer), and how policing decisions were made and actions taken based on those predictions (i.e., post-processing layer). A search of Newsbank's Access World News database revealed that news stories about LAPD's predictive policing technologies began to show up sporadically in the media and in LAPD press releases from 2012 to 2014 (e.g., see Refs. [105−107]) (2−3 years post-implementation), which may be why protests and activism against the technologies only began to gain traction around 2016 (5 years post-implementation). An organization known as "Stop LAPD Spying Coalition" requested public records about the LAPD's use of LASER in 2017 and 2018 due to concerns about unfair LAPD targeting of and forced interactions with Los Angeles residents, and then filed a lawsuit against the department in 2018 when the information was not provided[94]. As a result, the LAPD began to release many of the records to the public, albeit slowly[94].

The public is unlikely to support departments administering predictive policing technologies when they fail to provide transparent, thorough, and honest information on their benefits and risks across the end stages of the DSL. The black box effect during the post-processing stages of the LAPD's DSLs (i.e., decision-making/actions taken) fueled public mistrust specific to system predictions providing a covert excuse for racialized enforcement of the law and institutional racism[79, 100]. Predictive policing DSLs can enable "tech-washing" where communities and people of color are specifically, disproportionately, and (potentially) unjustly targeted with a façade of data-driven ethics and objectivity[108]. Through tech-washing, departments can operate absent sanctions or responsibility, which can reinforce harmful stereotypes and systemic injustice and facilitate their perpetuation. In this manner, the LAPD failed to meet its fiduciary responsibility to ensure that the DSL was not misused by police.

4.3 Facilitating fiduciary responsibility: Open-box DSLs

The LAPD's deployment of LASER and PredPol illustrates how a societal institution's failure to meet fiduciary responsibility[1] in its development and implementation of DSLs resulted in the public perceiving the department as lacking integrity—untrustworthy, dishonest, racist, and unjust toward people and communities of color. Department actions taken following DSL prediction output in the post-processing stages—hidden from stakeholders—are suspected to be immoral, discriminatory, and harmful to communities (e.g., disproportionate targeting and arrests of people of color, overpolicing of their communities, and worsening mass incarceration). We argue that police departments' development and deployment of predictive policing DSLs in secret lack accountability and have great potential to upset the functioning of society[1]. The LAPD's failure to build trust in their implementation of the predictive policing DSLs resulted in residents losing trust in, and support for, not only those DSLs but the LAPD overall[¤].

Such outcomes demonstrate the necessity of assessing and potentially altering the design, deployment, and usage of a DSL within the pre-processing, model building, and post-processing layers. Specifically, these outcomes show that DSLs must operate within an open-box regime, as we depict in Fig. 1. This regime provides the mechanisms for stakeholders to be confident that the institution is meeting its fiduciary responsibility through understanding interpretations of model outputs and thorough communication of the DSL's final decisions. Additionally, the open-box regime facilitates stakeholder and expert auditing of the overall DSL.

Below we present some specific examples of open-box DSL actions police departments could take as part of the last two stages of the post-processing layer to enhance fiduciary responsibility. In the "Communicate" stage, police leadership could provide frequent and descriptive/transparent press releases and social media posts that indicate the need for predictive policing, the data being used, how community members will be affected, process and status of design and implementation, and methods used to reduce bias. Similarly, departments could also hold media events or town halls to answer questions from the press and public, as well as take note of any concerns that should be considered prior to deploying resources/actions (i.e.,

[¤]It is important to note that many members of the public and police officers also support the usage of the LAPD's predictive policing technologies (e.g., Refs. [71, 96]).

decision-making) in impacted communities (i.e., last stage of the post-processing layer). In regards to the latter, police leadership should consider again communicating with the public about how such concerns were or will be addressed. Utilizing multiple methods of communication may have a stronger effect on public perceptions of the department's integrity and fiduciary responsibility.

As part of the "Deploy" stage, law enforcement may want to use a document similar to a contract (or create mandatory procedural guidelines) that explicates the actions they will take based on the output/prediction of the predictive policing DSL. This document could be updated periodically to reflect lessons learned (and subsequently alter or enhance communication content and strategies) and public feedback received across the duration of AI-based decision-making and deployment. We do not provide suggestions here for the "Interpret" stage of the DSL post-processing layer because this stage is internal to the police department and dependent on insider (i.e., police officer/leadership) knowledge.

5 Conclusion and Future Work

Facilitating public trust in AI and institutions deploying them is imperative for maintaining social cohesion. The recreancy theorem delineates three dimensions of public support for technical innovations. In this paper, we considered the dimension of fiduciary responsibility, which is the public's perception that a technology is designed and employed with integrity and honesty, performs as intended, does not create disparate impacts to vulnerable populations, and is not misused by agents. To formalize AI fiduciary responsibility, we described AI within the larger perspective of a DSL. The DSL framework provides multiple methods to precisely describe fiduciary responsibility of technologies affecting the public welfare and how institutions can meet their fiduciary responsibility. We investigated the example of the LAPD not meeting its fiduciary responsibility in its deployment of predictive policing technology.

We envision future work in at least two directions. First, the description of fiduciary responsibility within the DSL framework can be further refined and quantified. We have introduced multiple aspects and manifestations of fiduciary responsibility within DSLs, but did not consider system dynamics in our investigation—this consideration warrants a full, separate analysis given the complexity of dynamics involving DSLs. In addition, our work here was only qualitative in nature, and we did not propose mechanisms for institutions to quantify whether they have met their fiduciary responsibilities. Second, further development of the DSL framework can be more broadly utilized in a number of potential ways to facilitate public trust in AI through documentation, regulatory requirements, and a technologically-aware public. This, too, warrants a full, separate analysis from the one we have presented in this work.

The current paper uniquely situated sources/causes of mistrust in varying DSL contexts wherein each layer/stage has processes through which fiduciary responsibility can (or cannot) be addressed. We illustrated the importance (and necessity) of embedding fiduciary responsibility across the DSL workflow over time wherein institutions' decisions and deployment of actions in the post-processing layer can influence the public in profound and consequential ways.

Acknowledgment

This work was supported by the National Science Foundation and the National Geospatial Intelligence Agency (No. 1830254) and the National Science Foundation (No. 1934884).

References

[1] S. G. Sapp, S. Dorius, K. Bertelson, and S. Harper, Public support for government use of network surveillance: An empirical assessment of public understanding of ethics in science administration, *Public Understanding of Science*, vol. 31, no. 4, pp. 489–506, 2022.

[2] Y. Kao and S. G. Sapp, The effect of cultural values and institutional trust on public perceptions of government use of network surveillance, *Technology in Society*, vol. 70, p. 102047, 2022.

[3] S. G. Sapp, C. Arnot, J. Fallon, T. Fleck, D. Soorholtz, M. Sutton-Vermeulen, and J. J. H. Wilson, Consumer trust in the US food system: An examination of the recreancy theorem, *Rural Sociology*, vol. 74, no. 4, pp. 525–545, 2009.

[4] G. O. Mohler, M. B. Short, S. Malinowski, M. Johnson, G. E. Tita, A. L. Bertozzi, and P. J. Brantingham, Randomized controlled field trials of predictive policing, *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1399–1411, 2015.

[5] K. Lum and W. Isaac, To predict and serve? *Significance*, vol. 13, no. 5, pp. 14–19, 2016.

[6] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian, Runaway feedback loops in predictive policing, in *Proc. 1st Conference on Fairness,*

Accountability and Transparency, New York, NY, USA, 2018, pp. 160–171.

[7] M. Hengstler, E. Enkel, and S. Duelli, Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices, *Technological Forecasting and Social Change*, vol. 105, pp. 105–120, 2016.

[8] K. Siau and W. Wang, Building trust in artificial intelligence, machine learning, and robotics, *Cutter Business Technology Journal*, vol. 31, no. 2, pp. 47–53, 2018.

[9] R. Dobbe, S. Dean, T. Gilbert, and N. Kohli, A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics, arXiv preprint arXiv: 1807.00553, 2018.

[10] J. D. Lee and K. A. See, Trust in automation: Designing for appropriate reliance, *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[11] L. Jaume-Palasí and M. Spielkamp, Ethics and algorithmic processes for decision making and decision support, https://algorithmwatch.org/de/wp-content/uploads/2017/06/AlgorithmWatch_Working-Paper_No_2_Ethics_ADM.pdf, 2017.

[12] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, The ethics of algorithms: Mapping the debate, *Big Data & Society*, doi: 10.1177/2053951716679679.

[13] S. Barocas and A. D. Selbst, Big data's disparate impact, http://papers.ssrn.com/abstract=2477899, 2016.

[14] H. Mouzannar, M. I. Ohannessian, and N. Srebro, From fair decision making to social equality, in *Proc. Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, USA, 2019, pp. 359–368.

[15] B. Knowles and J. T. Richards, The sanction of authority: Promoting public trust in AI, in *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event, Canada, 2021, pp. 262–271.

[16] S. Biswas, M. Wardat, and H. Rajan, The art and practice of data science pipelines: A comprehensive study of data science pipelines in theory, in-the-small, and in-the-large, in *Proc. 2022 IEEE/ACM 44th International Conference on Software Engineering*, Pittsburgh, PA, USA, 2022, pp. 2091–2103.

[17] G. Nguyen, S. Dlugolinsky, M. Bobák, V. Tran, Á. L. García, I. Heredia, P. Malík, and L. Hluchý, Machine learning and deep learning frameworks and libraries for large-scale data mining: A survey, *Artificial Intelligence Review*, vol. 52, no. 1, pp. 77–124, 2019.

[18] R. S. Olson, N. Bartley, R. J. Urbanowicz, and J. H. Moore, Evaluation of a tree-based pipeline optimization tool for automating data science, in *Proc. Genetic and Evolutionary Computation Conference*, Denver, CO, USA, 2016, pp. 485–492.

[19] H. Wickham, Data science: How is it different to statistics? IMS Bulletin, https://imstat.org/2014/09/04/data-science-how-is-it-different-to-statistics%E2%80%89/, 2014.

[20] S. A. Hong and T. Hunter, Build, scale, and deploy deep learning pipelines with ease, https://www.databricks.com/blog/2017/09/06/build-scale-deploy-deep-learning-pipelines-ease.html, 2017.

[21] S. Todd and D. Dietrich, Computing resource re-provisioning during data analytic lifecycle, US Patent 9619550B1, 2017.

[22] R. Garcia, V. Sreekanti, N. Yadwadkar, D. Crankshaw, J. E. Gonzalez, and J. M. Hellerstein, Context: The missing piece in the machine learning lifecycle, https://rlnsanz.github.io/dat/Flor_CMI_18_CameraReady.pdf, 2018.

[23] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, Data lifecycle challenges in production machine learning: A survey, *ACM SIGMOD Record*, vol. 47, no. 2, pp. 17–28, 2018.

[24] L. Zhou, How to build a better machine learning pipeline, https://www.datanami.com/2018/09/05/how-to-build-a-better-machine-learning-pipeline/, 2018.

[25] J. M. Wing, The data life cycle, https://hdsr.mitpress.mit.edu/pub/577rq08d/release/3, 2018.

[26] R. Ashmore, R. Calinescu, and C. Paterson, Assuring the machine learning lifecycle: Desiderata, methods, and challenges, *ACM Computing Surveys*, vol. 54, no. 5, pp. 1–39, 2022.

[27] W. R. Freudenburg, Risk and recreancy: Weber, the division of labor, and the rationality of risk perceptions, *Social Forces*, vol. 71, no. 4, pp. 909–932, 1993.

[28] M. Alario and W. Freudenburg, The paradoxes of modernity: Scientific advances, environmental problems, and risks to the social fabric? *Sociological Forum*, vol. 18, no. 2, pp. 193–214, 2003.

[29] G. Roth and C. Wittich, *Economy and Society: An Outline of Interpretive Sociology*. Berkeley, CA, USA: University of California Press, 1978.

[30] J. A. Colquitt, B. A. Scott, and J. A. LePine, Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance, *Journal of Applied Psychology*, vol. 92, no. 4, pp. 909–927, 2007.

[31] K. Ball, S. D. Esposti, S. Dibb, V. Pavone, and E. Santiago-Gomez, Institutional trustworthiness and national security governance: Evidence from six European countries, *Governance*, vol. 32, no. 1, pp. 103–121, 2019.

[32] B. Barber, *The Logic and Limits of Trust*. New Brunswick, NJ, USA: Rutgers University Press, 1983.

[33] K. Blomqvist, Trust in a dynamic environment: Fast trust as a threshold condition for asymmetric technology partnership formation in the ICT sector, *Trust in Pressure: Investigations of Trust and Trust Building in Uncertain Circumstances*, doi: 10.4337/9781845427962.00011.

[34] K. Blomqvist, P. Hurmelinna, and R. Seppänen, Playing the collaboration game right—Balancing trust and contracting, *Technovation*, vol. 25, no. 5, pp. 497–504, 2005.

[35] M. Deutsch, Trust and suspicion, *Journal of Conflict Resolution*, vol. 2, no. 4, pp. 265–279, 1958.

[36] S. G. Sapp and T. Downing-Matibag, Consumer acceptance of food irradiation: A test of the recreancy theorem, *International Journal of Consumer Studies*, vol. 33, no. 4, pp. 417–424, 2009.

[37] S. Kim and J. Lee, E-participation, transparency, and trust in local government, *Public Administration Review*, vol. 72, no. 6, pp. 819–828, 2012.
[38] R. Hardin, Conceptions and explanations of trust, in *Trust in Society*, K. Cook, ed. New York, NY, USA: Russell Sage Foundation, 2001, pp. 3–39.
[39] T. C. Earle and G. Cvetkovich, *Social Trust: Toward a Cosmopolitan Society*. Westport, CT, USA: Greenwood Publishing Group, 1995.
[40] M. Siegrist and G. Cvetkovich, Perception of hazards: The role of social trust and knowledge, *Risk analysis*, vol. 20, no. 5, pp. 713–720, 2000.
[41] C. A. Cooper, H. G. Knotts, and K. M. Brennan, The importance of trust in government for public administration: The case of zoning, *Public Administration Review*, vol. 68, no. 3, pp. 459–468, 2008.
[42] F. D. Schoorman, R. C. Mayer, and J. H. Davis, An integrative model of organizational trust: Past, present, and future, *Academy of Management Review*, vol. 32, pp. 344–354, 2007.
[43] S. G. Sapp, P. F. Korsching, C. Arnot, and J. J. H. Wilson, Science communication and the rationality of public opinion formation, *Science Communication*, vol. 35, no. 6, pp. 734–757, 2013.
[44] J. Anderson, L. Rainie, and A. Luchsinger, Artificial intelligence and the future of humans, *Pew Research Center*, vol. 10, p. 12, 2018.
[45] P. Hitlin and L. Rainie, Facebook algorithms and personal data, https://www.pewresearch.org/internet/2019/01/16/facebook-algorithms-and-personal-data/, 2019.
[46] K. Olmstead and A. Smith, Americans and cybersecurity, https://assets.pewresearch.org/wp-content/uploads/sites/14/2017/01/26102016/Americans-and-Cyber-Security-final.pdf, 2017.
[47] L. Rainie, J. Anderson, and D. Page, Code-dependent: Pros and cons of the algorithm age, *Pew Research Center*, 2017.
[48] L. Rainie and J. Anderson, The internet of things connectivity binge: What are the implications? *Pew Research Center*, 2017.
[49] J. D. Lewis and A. Weigert, Trust as a social reality, *Social Forces*, vol. 63, no. 4, pp. 967–985, 1985.
[50] G. J. Hofstede, Intrinsic and enforceable trust: A research agenda, presented at 99th EAAE Seminar, Bonn, Germany, 2006.
[51] M. Wißner, S. Hammer, E. Kurdyukova, and E. André, Trust-based decision-making for the adaptation of public displays in changing social contexts, *Journal of Trust Management*, vol. 1, p. 6, 2014.
[52] IBM Research AI, Trustworthy AI, https://research.ibm.com/topics/trustworthy-ai, 2020.
[53] European Commission, White paper: On artificial intelligence–a European approach to excellence and trust, https://ec.europa.eu/futurium/en/system/files/ged/white_paper_ai_19_02_2020.pdf, 2020.
[54] Z. Yan and S. Holtmanns, Trust modeling and management: From social trust to digital trust, in *Computer Security, Privacy and Politics: Current Issues, Challenges and Solutions*, R. Subramanian, ed. Hershey, PA, USA: IGI Global, 2008, pp. 290–323.
[55] R. Harper, *Trust, Computing, and Society*. Cambridge, UK: Cambridge University Press, 2014.
[56] A. Jobin, M. Ienca, and E. Vayena, The global landscape of AI ethics guidelines, *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.
[57] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. V. Moorsel, The relationship between trust in AI and trustworthy machine learning technologies, in *Proc. 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 2020, pp. 272–283.
[58] M. Arnold, R. K. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski, et al., FactSheets: Increasing trust in AI services through supplier's declarations of conformity, *IBM Journal of Research and Development*, vol. 63, nos. 4&5, pp. 1–13, 2019.
[59] A. Jacovi, A. Mojsilović, T. Miller, and Y. Goldberg, Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI, in *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event, Canada, 2021, pp. 624–635.
[60] R. C. Mayer, J. H. Davis, and F. D. Schoorman, An integrative model of organizational trust, *Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995.
[61] K. Hawley, Trust, distrust and commitment, *Noûs*, vol. 48, no. 1, pp. 1–20, 2014.
[62] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144.
[63] B. Mittelstadt, C. Russell, and S. Wachter, Explaining explanations in AI, in *Proc. Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, USA, 2019, pp. 279–288.
[64] J. Richards, D. Piorkowski, M. Hind, S. Houde, and A. Mojsilovic, A methodology for creating AI factsheets, arXiv preprint arXiv: 2006.13796, 2020
[65] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilovic, R. Nair, K. N. Ramamurthy, D. Reimer, A. Olteanu, et al., FactSheets: Increasing trust in AI services through supplier's declarations of conformity, arXiv preprint arXiv: 1808.07261, 2018.
[66] F. D. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology, *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989.
[67] F. D. Davis, User acceptance of information technology: System characteristics, user perceptions and behavioral impacts, *International Journal of Man-Machine Studies*, vol. 38, no. 3, pp. 475–487, 1993.
[68] S. Wachter, B. Mittelstadt, and C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, *Harv. JL & Tech.*, vol. 31, no. 2, p. 841, 2018.
[69] L. Gilpin, A. Paley, M. Alam, S. Spurlock, and K. Hammond, "Explanation" is not a technical term: The problem of ambiguity in XAI, arXiv preprint arXiv: 2207.00007, 2022.
[70] N. Kohli, R. Barreto, and J. A. Kroll, Translation tutorial:

A shared lexicon for research and practice in human-centered software systems, in *Proc. 1st Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 2018, p. 7.

[71] T. Lau, Predictive policing explained, https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained, 2020.

[72] S. Thomson, 'predictive policing': Law enforcement revolution or just new spin on old biases? Depends who you ask, https://www.cbc.ca/news/world/crime-los-angeles-predictive-policing-algorithms-1.4826030, 2018.

[73] R. B. Santos, Predictive policing: Where's the evidence? in *Police Innovation*, D. Weisburd and A. A. Braga, eds. Cambridge, UK: Cambridge University Press, 2019, pp. 366–396.

[74] W. Hardyns and A. Rummens, Predictive policing as a new tool for law enforcement? Recent developments and challenges,, *European Journal on Criminal Policy and Research*, vol. 24, no. 3, pp. 201–218, 2018.

[75] A. Meijer and M. Wessels, Predictive policing: Review of benefits and drawbacks, *International Journal of Public Administration*, vol. 42, no. 12, pp. 1031–1039, 2019.

[76] E. A. Carson and D. Golinelli, Prisoners in 2012: Trends in admissions and releases, 1991–2012, https://bjs.ojp.gov/content/pub/pdf/p12tar9112.pdf, 2014.

[77] A. Rosenberg, A. K. Groves, and K. M. Blankenship, Comparing black and white drug offenders: Implications for racial disparities in criminal justice and reentry policy and programming, *Journal of Drug Issues*, vol. 47, no. 1, pp. 132–142, 2017.

[78] M. Alexander, The new Jim crow, *Ohio St. J. Crim. L.*, vol. 9, no. 1, pp. 7–26, 2011.

[79] R. Richardson, J. M. Schultz, and K. Crawford, Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice, *NYUL Rev. Online*, vol. 94, p. 15, 2019.

[80] A. J. Ritchie, *Invisible No More: Police Violence Against Black Women and Women of Color*. Boston, MA, USA: Beacon Press, 2017.

[81] American Civil Liberties Union, Statement of concern about predictive policing by ACLU and 16 civil rights privacy, racial justice, and technology organizations, https://www.aclu.org/other/statement-concern-about-predictive-policing-aclu-and-16-civil-rights-privacy-racial-justice, 2016.

[82] L. Barrett, Reasonably suspicious algorithms: Predictive policing at the United States border, *NYU Rev. L. & Soc. Change*, vol. 41, p. 327, 2017.

[83] A. Edwards, Big data, predictive machines and security: The minority report, in *the Routledge Handbook of Technology, Crime and Justice*, M. R. McGuire and T. J. Holt, eds. Oxfordshire, UK: Routledge, 2017, pp. 451–461.

[84] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY, USA: Crown Books, 2016.

[85] J. Bhuiyan, LAPD ended predictive policing programs amid public outcry. A new effort shares many of their flaws, the Guardian, https://www.theguardian.com/us-news/2021/nov/07/lapd-predictive-policing-surveillance-reform, 2021.

[86] E. R. Moravec, Do algorithms have a place in policing? the Atlantic, https://www.theatlantic.com/politics/archive/2019/09/do-algorithms-have-place-policing/596851/, 2019.

[87] A. G. Ferguson, Predictive policing and reasonable suspicion, *Emory Law Journal*, vol. 62, p. 259, 2012.

[88] A. Tarantola, Predictive policing' could amplify today's law enforcement issues, Engadget, https://www.engadget.com/predictive-policing-privacy-civil-rights-dangers-133040971.html, 2020.

[89] A. G. Ferguson, *The Rise of Big Data Policing*. New York, NY, USA: New York University Press, 2017.

[90] T. Aougab, F. Ardila, J. Athreya, E. Goins, and C. Hoffman, Letters to the editor: Boycott collaboration with police, *Notices Amer. Math. Soc.*, vol. 67, no. 9, p. 1293, 2020.

[91] W. D. Heaven, Predictive policing algorithms are racist. They need to be dismantled, MIT Technology Review, https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/, 2020.

[92] M. Hvistendahl, How the LAPD and Palantir use data to justify racist policing, the Intercept, https://theintercept.com/2021/01/30/lapd-palantir-data-driven-policing/, 2021.

[93] C. Linder, Why hundreds of mathematicians are boycotting predictive policing, Popular Mechanics, https://www.popularmechanics.com/science/math/a32957375/mathematicians-boycott-predictive-policing/, 2020.

[94] G. Baek and T. Mooney, LAPD not giving up on data-driven policing, even after scrapping controversial program, https://www.cbsnews.com/news/los-angeles-police-department-laser-data-driven-policing-racial-profiling-2-0-cbsn-originals-documentary/, 2020.

[95] I. Ayres and J. Borowsky, Study of racially disparate outcomes in the Los Angeles Police Department, ACLU of Southern California, https://www.aclusocal.org/sites/default/files/wp-content/uploads/2015/09/11837125-LAPD-Racial-Profiling-Report-ACLU.pdf, 2008.

[96] I. Lapowski, How the LAPD uses data to predict crime, https://www.wired.com/story/los-angeles-police-department-predictive-policing/, 2018.

[97] M. P. Smith, Review of selected Los Angeles Police Department data-driven policing strategies, Los Angeles Police Commission, https://www.lapdpolicecom.lacity.org/031219/BPC_19-0072.pdf, 2019.

[98] M. Tonkin, J. Woodhams, R. Bull, J. W. Bond, and E. J. Palmer, Linking different types of crime using geographical and temporal proximity, *Criminal Justice and Behavior*, vol. 38, no. 11, pp. 1069–1088, 2011.

[99] F. Yang, Predictive policing, *Oxford Research Encyclopedia of Criminology and Criminal Justice*, doi: 10.1093/acrefore/9780190264079.013.508.

[100] T. Mooney and G. Baek, Is artificial intelligence making racial profiling worse? https://www.cbsnews.com/news/artificial-intelligence-racial-profiling-2-0-cbsn-originals-documentary/, 2020.

[101] E. Bakke, Predictive policing: The argument for public

transparency, *NYU Ann. Surv. Am. L.*, vol. 74, pp. 131–171, 2018.
[102] C. Chang, LAPD officials defend predictive policing as some groups call for its end, https://www.police1.com/patrol-issues/articles/lapd-officials-defend-predictive-policing-as-some-groups-call-for-its-end-PNfxLd2b6JajAZDs/, 2018.
[103] Stop LAPD Spying Coalition, Before the bullet hits the body: Dismantling predictive policing in Los Angeles, https://stoplapdspying.org/wp-content/uploads/2018/05/Before-the-Bullet-Hits-the-Body-May-8-2018.pdf, 2018.
[104] L. Miller, LAPD will end controversial program that aimed to predict where crimes would occur, https://www.latimes.com/california/story/2020-04-21/lapd-ends-predictive-policing-program, 2020.
[105] R. Bailey, Stopping crime before it starts, Reason, https://reason.com/2012/07/10/predictive-policing-criminals-crime/, 2012.
[106] Los Angeles Police Department, LAPD foothill community police station announces "international predpol day of action", https://www.lapdonline.org/newsroom/lapd-foothill-community-police-station-announces-international-predpol-day-of-action-na13136bb/, 2013.
[107] A. C. Madrigal, Toward a complex, realistic, and moral tech criticism, the Atlantic, https://www.theatlantic.com/technology/archive/2013/03/toward-a-complex-realistic-and-moral-tech-criticism/273996/, 2013.
[108] S. Egbert and M. Mann, Discrimination in predictive policing: The (dangerous) myth of impartiality and the need for STS analysis, in *Automating Crime Prevention, Surveillance, and Military Operations*, A. Završnik and V. Badalič, eds. Cham, Switzerland: Springer, 2021, pp. 25–46.

Eric S. Weber received the PhD degree in mathematics from University of Colorado. He is a professor and chair of mathematics at Iowa State University. His research interests include harmonic analysis and approximation theory. His past research includes developing novel wavelet transforms for image processing and reproducing kernel methods for the harmonic analysis of fractals. His current research projects include the development of new algorithms for processing distributed spatiotemporal datasets in order to increase understanding of human dynamics; extending alternating projection methods for optimization in non-Euclidean geometries; using harmonic analysis techniques for understanding the approximation properties of neural networks; and developing machine learning techniques to improve the diagnosis of severe wind occurrences.

Shannon B. Harper is an assistant professor of criminal justice in the Department of Sociology and Criminal Justice, Iowa State University. Dr. Harper's research explores the relationship between intimate partner violence and homicide, as well as how crime victims perceive the usefulness and accessibility of institutions that provide victim services, including the criminal justice system. Such scholarship includes focus on police decision-making processes that influence their interactions with victims and the public at large. Dr. Harper's work is published in numerous highly ranked criminological/sociological journals, including the *Journal of Interpersonal Violence*, *Public Understanding of Science*, *Feminist Criminology*, and the *American Journal of Criminal Justice*.