AI & SOCIETY
https://doi.org/10.1007/s00146-022-01417-y

OPEN FORUM

Separating facts and evaluation: motivation, account, and learnings from a novel approach to evaluating the human impacts of machine learning

Ryan Jenkins[1] · **Kristian Hammond**[2] · **Sarah Spurlock**[2] · **Leilani Gilpin**[3]

Received: 30 May 2021 / Accepted: 21 February 2022

Abstract
In this paper, we outline a new method for evaluating the human impact of machine-learning (ML) applications. In partnership with Underwriters Laboratories Inc., we have developed a framework to evaluate the impacts of a particular use of machine learning that is based on the goals and values of the domain in which that application is deployed. By examining the use of artificial intelligence (AI) in particular domains, such as journalism, criminal justice, or law, we can develop more nuanced and practically relevant understandings of key ethical guidelines for artificial intelligence. By decoupling the extraction of the facts of the matter from the evaluation of the impact of the resulting systems, we create a framework for the process of assessing impact that has two distinctly different phases.

Keywords Machine learning · Impact assessment · Operationalizing · Practice dependence · Design for values

1 Introduction

The computer science industry has come to appreciate the significance of anticipating the ethically significant effects of machine learning. Academic and popular literature outlining the various concerns with machine learning have proliferated, and dozens of companies, nonprofits, governmental organizations, and other entities have promulgated codes of ethics to guide technologists in developing their models (e.g., IEEE 2018; Fjeld et al. 2020). While we celebrate the broad agreement over the fundamental ethical dimensions of artificial intelligence—fairness, accountability, transparency, and explainability—many, including the authors, have also grown despondent that the conversation has come to rest around principles that are vague. For example, the nature and significance of "fairness" in AI differs profoundly from one use context to another.

In this paper, we outline a new method for the evaluation of the human impact of machine learning. In partnership with Underwriters Laboratories Inc., we have developed a framework to evaluate the impacts of a particular use of machine learning that is *based on the goals and values of the social domain in which that application is deployed*. Our hope is to move beyond the top-down approach in AI ethics that has reigned (Allen et al. 2005), and to move to a middle-out system instead. By examining the use of AI in particular domains, such as journalism, criminal justice, or law, we can develop more nuanced and practically relevant understandings of key ethical guidelines for artificial intelligence, and do so in a way that earns the buy-in from the very practitioners whose subject matter expertise matters most. This part of the approach is *middle-down*. At the same time, we can search for novel overarching principles that connect the concerns of multiple domains. This part of the approach is *middle-up*.

✉ Ryan Jenkins
ryjenkin@calpoly.edu

Kristian Hammond
kristian.hammond@northwestern.edu

Sarah Spurlock
sarah.spurlock@northwestern.edu

Leilani Gilpin
lgilpin@ucsc.edu

[1] Philosophy Department, California Polytechnic State University, San Luis Obispo, USA

[2] Computer Science Department, Northwestern University, Evanston, USA

[3] Computer Science and Engineering Department, University of California, Santa Cruz, USA

Published online: 16 March 2022

Our approach is based on the idea of decoupling *facts* from *evaluation*. By decoupling the extraction of the facts of the matter from the evaluation of the impact of the resulting systems, we create a framework for the process of assessing impact that has two distinct phases. For systems based on models developed through machine learning, the facts include issues of data (acquisition, cleaning, normalization, coverage, etc.), algorithmic choice and development (algorithms used, feature selection, performance expectation, training and testing), and system interaction (what function is the system playing in decision making, how are results framed, and how is performance measured). The evaluation of the impact of systems based on ML models becomes one of determining how the facts of a system relate to the goals and values that it is designed to support.

In Sect. 2, we discuss our motivation for developing this approach as filling a gap in the literature on approaches to artificial intelligence ethics. In Sect. 3, we provide an outline of the framework in its current form, including a set of heuristic questions to use at each stage of analysis. In Sect. 4, we discuss the concrete methods that we used to stress test this framework during workshops in November 2020 and June 2021 and discuss our major findings so far. In Sect. 5, we discuss next steps in refining the framework.

2 Motivation

Efforts to guide the ethical development of artificial intelligence have proliferated at an astonishing pace over the last few years—96 since 2018 alone (Zhang et al. 2021: 130). While these guidelines have been produced by governments, professional organizations, NGOs, private corporations, universities, think tanks, and others, they are largely top-down (Allen et al. 2005): they put forth general principles at the highest level which can arch over the development and deployment of artificial intelligence. Common themes among these are fairness, accountability, transparency, explainability (FATE); justice, human rights, and so on. This discussion has spawned additional work across disciplines to investigate the nature of these values and operationalize them. With the introduction of the European Union's (EU's) General Data Protection Regulation (GDPR), for example, which guarantees a citizen's right to an explanation (Goodman and Flaxman 2017), there has been a flurry of activity exploring the nature of explanation and the nature of this purported duty towards data subjects (Kaminski 2019; Selbst and Powles 2018).

As the development and refinement of AI techniques continues, identifying these overarching values and investigating their nature is clearly important. But it comes with several costs which are also becoming more appreciated. First, what these frameworks boast in generality they sacrifice in power and action-guidingness. We are not the only ones to share this view. See, for example, Zhang et al. (2021), who bemoan that "the vague and abstract nature of those principles fails to offer direction on how to implement AI-related ethics guidelines" (129). Mittelstadt (2019) notes that the field of AI ethics crucially lacks "proven methods to translate principles into practice" and in particular the "professional history and norms" (1) that could furnish more concrete guidance—precisely what we hope to scaffold with this project.

For one thing, these abstract principles require a tremendous amount of work to operationalize, and they have led to disagreements at the technical level around what measures of success might be appropriate for judging, for example, the fairness of a model (Alikhademi et al. 2021) or its accountability (Wieringa 2020). Canca (2020), for example, belabors the point that operationalizing AI principles will differ across domains, and the meaning of specific values will change across contexts, mentioning specifically how the value of *transparency* might differ across law enforcement, sustainability, and medical applications (20–21). Similarly, Madaio et al (2020) adopt an iterative process to operationalize *fairness* which, they note, depends significantly on the context of the specific application.

Second, these approaches are also ignorant of the context of particular deployments. This is just what it is to say that these principles are *maximally general*; in fact, we believe they are general to a fault. The fact that they are agnostic about the domains in which artificial intelligence is deployed is an obstacle to their operationalization. There are significant and reasonable disagreements between practitioners in different domains about the nature of the FATE and other concepts. Similarly, the *importance* of each of these considerations might differ from one domain to another. If a model is opaque (i.e., not transparent), this might be unproblematic if the model is used to recommend ads to a user—but this could be a conclusive reason to reject its use in the context of banking.

On the other end of the spectrum, there is a vast and growing literature that examines and critiques specific instances of the deployment of artificial intelligence. This "bottom-up" work is also important, but it suffers from weaknesses that are the inverse of the top-down approach. This literature provides some of the most precise critiques and useful action-guidance; but the utility of these insights is limited because they are not portable. For example, the special issue introduced by Rodgers (2021) explores issues related to using AI in advertising, including optimizing ad placement, in-store advertising experiences, and using AI influencers. These guidelines are difficult to generalize to similar fields since the goals and values of advertising are different from those of, say, journalism or technical writing—nearby domains which have no need of in-store experiences and

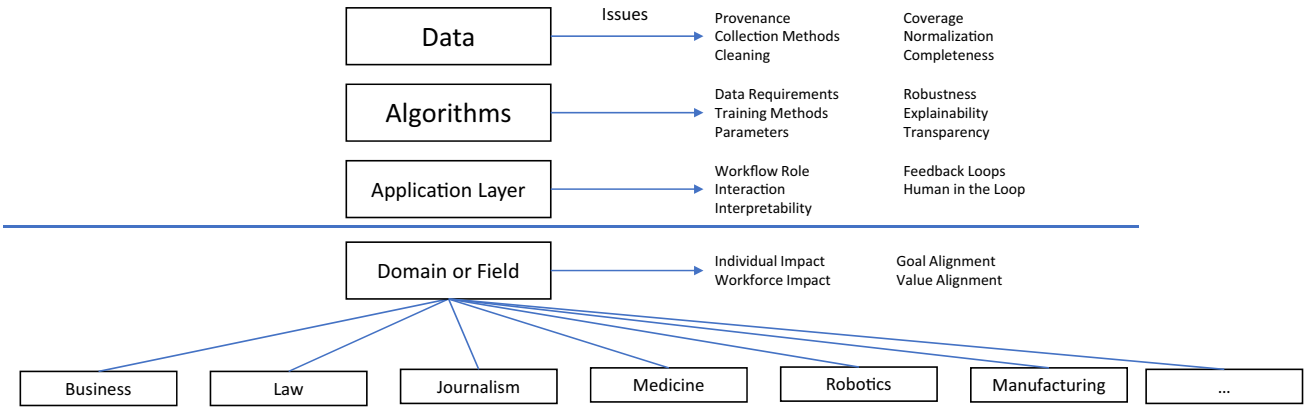

Fig. 1 Our framework for the evaluation of the impact of systems based on machine learning

which would frown upon the idea of influencers. Brey (2004) examines CCTV with facial recognition built in and identifies function creep, privacy, and error as potential problems. But the value and importance of privacy might differ when this technology is deployed in different domains, i.e., policing versus military, or during large sporting events, and this analysis ignores those domain distinctions by focusing simply on "public and semipublic places" (97). In addition, discussing facial recognition, Selinger and Leong (2021) note several times that their conclusions will vary between contexts such as law enforcement, employment, or marketing, suggesting that porting their insights to those domains will require additional work. Critiquing the use of AI for finance, Max et al (2020) identify, among other issues, "the individualization of service offerings" (578). But notice this issue is more pertinent in finance, where individualization risks price discrimination—while this same feature might be good in another domain. In short, it is difficult to generalize these findings for AI broadly or to other domains. Finally, it would be prohibitively onerous to examine the impacts of every instance of machine learning in society.

Spurred by these observations, we attempt to thread the needle by developing what we characterize as a *middle-out approach*. This approach takes *social domains* as the appropriate level of analysis, identifies the individual goals and values of those domains, and then explores how particular implementations of machine learning are liable to interact with those goals and values to produce positive or negative human impacts. Our approach seeks to balance the benefits of both generality and action-guidance while acknowledging the context-sensitivity of different values in AI ethics.

Of course, a full evaluation of the human impacts of machine-learning systems ought to include some reference to their broader context, since AI systems are but one part of a sociotechnical system (see van de Poel 2020; Kroes et al. 2006). The ultimate consequences of ML systems will be the outcomes of the interactions between human behavior, AI systems, and the norms of the domains in which they are embedded. This underscores the importance of working at the level of domains, since those provide natural boundaries for evaluating the impacts of a system as an outcome of its embedded use.

3 Outline of the framework

Our framework for the evaluation of the impact of systems based on machine learning divides the task into two phases, illustrated in Fig. 1. First, determining the design and development characteristics of applications and then the subsequent examination of how, given those facts, the system impacts the goals and values associated with a particular domain or field of use. This division allows us to establish the facts using agreed upon methods before confronting the evaluative phase that deals with issues where there is far less immediate agreement. This division also provides clarity as to what the conversation is about, where areas of agreement exist, and reduces the complexity of the problem.

We acknowledge that the distinction between facts and values in technical systems is controversial (Van de Poel 2015, 2020; Verbeek 2005). We agree that factual decisions about *how to design a system* are often value-laden, motivated by subjective preferences, and can have ethically significant implications (Tatum 1997). We wish only to separate the project of describing the technical details of a system from the project of evaluating the system, since those projects rely on different methods and expertise, and engage with different audiences. None of this is to rule out the possibility—indeed, the likelihood, which we discuss below—that decisions about the technical details of a machine-learning application will have evaluatively significant implications.

Finally, we underscore our view that the application of this framework should be part of an ongoing conversation

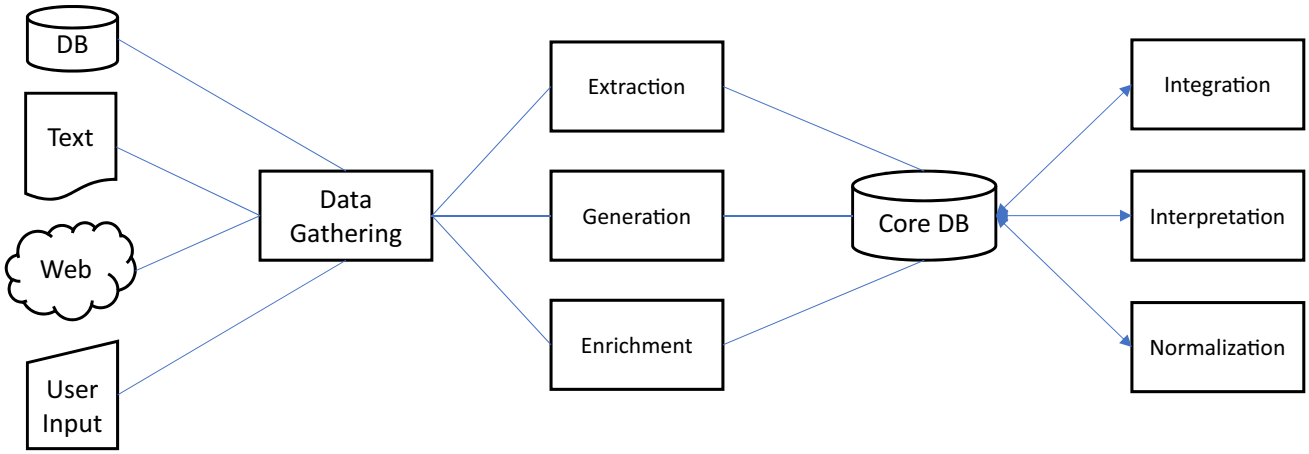

Fig. 2 Elements of the data collection and integration process

with developers and stakeholders, rather than a one-off event, and that feedback from those impacted by these applications should be continually revisited and incorporated, where appropriate, into the evaluation of the system.

3.1 Feature assessment and information gathering

The features of a ML-based application and its development can be broken into three classes of design decisions: (1) the nature of the data, how they were sourced and what processes were used to gather and integrate them; (2) the algorithms utilized in the learning process itself, the data requirements of the different approaches, developer feature selection, training methods, and approaches to testing and validation; (3) the ways in which the system was designed to interact with users and their relationship with the recommendations, assessments, predictions, or guidance that the system provides.

The focus on factors flowing from design decisions is based on two goals. First, as developers are building machine-learning systems, the focus on these design decisions provides them with the guidance they need to avoid building and deploying problematic systems from the beginning. Second, once an application is built and deployed, we direct the attention of those applying the framework to these design decisions as characteristics that could impact the evaluation phase. It is crucial for both groups to appreciate that the choice of the features of the data sets that drive learning can impact outcomes, e.g., see Amazon's use of performance review data that was itself gender biased (BBC 2018), or that population coverage in a data set skews performance, e.g., when all white faces are used to train a system approving passport photos (Barocas and Selbst 2016). Focusing on these decisions provides direction to both developers and evaluators as to what characteristics are important in terms of their potential impacts.

3.2 Data

Machine-learning systems are built on data and rely on data in their ongoing use. The facts of a system's data set are defined by the processes that were used to gather, clean, enhance, and integrate often multiple sources and data types. Each of the processes that make up the ecosystem of data collection and integration needs to be examined for issues of quality, completeness, and coverage. Each of the steps in the process could impact the performance of any system that is built on top of the models they produce. These steps are illustrated in Fig. 2.

In looking at the *original sources*, questions of their reliability, coverage, and user incentives that might bias the data have to be answered. These questions are aimed at uncovering features such as input forms that nudge users towards certain responses, the distribution and completeness of examples, and the validity of the sources themselves.

The *gathering* process needs to be looked at through a different lens. Examining how data was gathered, we consider whether the process itself introduces any issues that compromise completeness or coverage. Are there features of a data set that are left out or examples that are beyond the scope of the ingestion process? Are different data sources ingested in different ways or have their features been extracted in a manner that pulls them out of alignment?

As we consider how an initial data set is processed and enhanced, we need to go through a similar set of queries focusing on how data elements are *extracted* (e.g., pulling entries from text), *enriched* (mapping ambiguous pieces of text onto controlled vocabularies), and if new synthetic data sets have been *generated* to help support the learning that it will drive.

Finally, we have to consider the process of *integration* and examine how different data sets are brought together and if there are any places where the *interpretation* of the

data (i.e., fitting it to an existing ontology) or *normalization* of elements to serve integration might be introducing errors.

These inquiries into the sources and processing of the data are aimed at surfacing the sources of possible problems so that they can be identified and, if warranted by their impact, remedied.

3.3 Algorithms

Independent of the data are questions related to the algorithms supporting the learning process. In viewing algorithmic issues, the focus is on the two elements: (1) decisions related to the inputs and training itself and (2) those related to the features and expectations of the models that result. The first of these involves the choices that were made by developers that shape the model they produce. The second involves the model itself and its performance.

Developer choices include the features that are used by the system, the cycle of training and testing, the choice of specific algorithm and the training and retraining dynamics. Each of these issues impacts not just the level of performance of a system but also the nature of problems that might arise using it. Feature choices determine the characteristics that will define a credit assessment, performance review, diagnosis, etc. Training and testing choices can impact a model's coverage, skewing results even when the core data is balanced. In addition, choices about how a system is updated and retrained can create self-re-enforcing predictions.

Looking at the resulting models, we need to consider issues such as their levels of accuracy and their levels of opacity. Some systems provide more of a window into their operations and the features that they are utilizing than others. The different levels of transparency impact how and when different models can be utilized.

Different deployments and domains have different requirements. Fitting a single ML approach to all of them makes little sense and over-constrains the application of powerful technologies. To make decisions as to the applicability of technologies in specific situations with specific needs, we can extract characteristics such as levels of transparency that can later be used to assess that applicability.

3.4 Application layer

The data and algorithms that make use of them result in models that can be viewed and tested independently of their utilization. To understand how these systems impact human health and safety, we have to also consider how they are used. To explore this, we need to ask questions aimed at uncovering the ways in which models are utilized in decision making once they are deployed.

What is the core functionality of the system? What does it do (e.g., categorization, recommendation, decision support, prediction, diagnosis, etc.)? Who are the users and what are their skills? Are they equipped to judge the outputs of the model or is there a danger of overtrusting those outputs (Kirkpatrick et al. 2017)? In addition, what is the role of the user in the system?

The goal here is not to determine whether the interactions are appropriate but to understand exactly what they are. That a hand handoff occurs from machine to user when the machine is unable to make a decision is a fact. A handoff may be appropriate for a system putting together a credit assessment but could lead to problems if the system is driving a car. Again, the details of the domain matter, and come into play in the next stage.

3.5 Evaluation

The facts related to the data, the algorithm, and the application layer serve as inputs into the process that can be used to evaluate the system. At this stage, evaluation must be done within the context of the *domain* in which this application will be situated. Analyzing an algorithm against the "goals and values" of a domain, we believe, is a novel and powerful approach, and equips us with a new set of analytical tools to judge the deployment of machine learning, and to more precisely articulate the trade-offs, benefits and drawbacks of its human impact.

We will take a moment to explain our theory of "domains," their nature, and their connection to what we are calling *goals* and *values*. Our theory is inspired by the neo-Aristotelian conception of *domains of practice*—which we call *domains* for short[1]—which is popular among contemporary neo-Aristotelians such as Walzer (2008) and MacIntyre (1981, 1988). The discussion of domains of practice also borrows heavily from the work done on *practice-dependence*[2], especially within the global justice debate, has benefited from contributions from philosophers of law, such as Dworkin (1986: ch. 2), and the global justice literature that followed in the wake of Rawls (James 2005).[3] Nissenbaum's

[1] In this and adjacent literatures, the definitional question remains contested and chaotic. In the works of neo-Aristotelians like Walzer, the words "practice," "domain," "sphere," and so on are used almost interchangeably, and they may include multiple kinds of institutions, organizations, and associations. After more than 100 years of scholarly study on institutions, Rhodes et al. note in their preface to the Oxford Handbook of Political Institutions that there is still no singular definition of an "institution" that enjoys widespread agreement (2006: xiii).

[2] See especially James (2005), whose paper was an important catalyst for the recent flurry of scholarship on practice-dependence. See also Jubb (2016), Erman and Möller (2015, 2016).

[3] Note that discussion and employment of this method abounds in adjacent fields. This includes the political science literature on the origin and nature of institutions, for those who champion a sociological approach to understanding institutions. Crespo (2016) is particu-

influential "contextual integrity" view is inspired by similar sensibilities, namely, examining existing social practices and generating "regulative ideals" appropriate to them based on the nature of the activity itself (2004, 2011).

We suggest that domains are collections of participants who consciously share a common enterprise and at least rough agreement on its purposes. This is inspired by Rawls' definition of a practice as "any form of activity specified by a system of rules which defines offices, roles, moves, penalties, defenses, and so on, and which gives the activity its structure," though less strict (1955: 3).[4] Borrowing from the literature on political institutions, we would compare our view of domains to the "sociological" view that institutions are exogenous, "woven into traditions, culture, norms, and preferences," embedding history and practices, and "more likely to be evolved than created" (Rhodes et al. 2006: xiii; see also Rhodes et al. 2006: xv). What Rhodes et al. (2006) categorize as the "sociological view of institutions" fits with our view of domains, for example: "institutions can be considered as embedding rules and routines that define what constitutes appropriate action… individuals are said to behave according to their sense of duty and obligation as structured by prevailing rules and routines" (xvi).[5]

A few modifications to Rawls' and Rhodes et al.' discussions will bring us closer to what we mean by domains. First, in weakening the concept of "rules" at play to something more such as *norms* or *guidelines*, the concept becomes capacious enough to include activities such as journalism, criminal justice, advertising, and medicine. Journalism does not have "rules" in the same way that a parliament has rules of order, but journalists certainly hew to norms or guidelines in pursuit of their goal. Second, as we explore below, the shared understanding of this common enterprise naturally generates shared norms of behavior that govern which actions are fitting or appropriate for practitioners. For example, it is widely frowned upon among journalists for them to pay their sources, because it could undermine their ability to report the truth.[6] In more mature professions, "a rule is a habit that has become normative, can be codified, and has been adopted by a group of people" (Hodgson 2006: 6; Crespo 2016: 881). Professional codes of ethics, which we mention below, are a perfect example of such codified, collectively accepted, normatively loaded habits.

This method seeks to excavate from concrete practices the implicit principles that are already in force among serious practitioners within a domain. The practitioners of a domain are specially placed, given their own behavior and that of those around them, to negotiate an equilibrium that satisfies the relevant goals and the concerns of the participants. One benefit of our approach is that the norms we infer through a close study of specific practices are more likely to be implementable and to be justifiable to the practitioners within a domain and the larger public that is affected by it.[7]

3.5.1 Goals and values

First, we try to understand the goal of a domain. This is accomplished in conversation with the practitioners in the domain, understanding what they take themselves to be doing and how it may be different from other adjacent domains—what is it that separates journalism from propaganda, for example? The goal of a domain is the contribution it makes to society or what those inside the domain are trying to accomplish. Much like specifying requirements during the standard engineering process, we suggest viewing goals as *moralized requirements* that must be met for a system to be acceptable (Van de Poel 2013; Richardson 1997).

We define a goal as "an outcome we hope to accomplish in a domain." When we use the word, "goal," we mean it *aspirationally* as opposed to *descriptively*. We are not trying to describe people's actual motivations, because they might be motivated by fame, reputation, money, vengeance or any number of less savory things that ought not guide the development of AI. We are interested in augmenting and catalyzing the positive contributions that these domains make to society. We developed several prompts to help practitioners

Footnote 3 (continued)

larly helpful in surveying theories of institutions and their consonance with the Aristotelian view we are sympathetic to here. According to Crespo, this Aristotelian view is consistent with the three reigning sociological theories of institutions, but fits best with the constitutive rule theory promoted by Crespo (2016: 868; and see Searle 2005). See also Hendriks and Guala (2015), which sparked renewed interest in the sociology of institutions. A further cluster of related research is found in the sociology of professions and professionalization.

[4] See also Lamarque (2010) on the Wittgensteinian precursors to Rawls' own suggestions about the nature of a practice.

[5] See also Hodgson, who says institutions are "integrated systems of rules that structure social interactions" (2015: 501), and elsewhere they are "established and prevalent social rules" (2006: 2).

[6] Some of these norms might be stringent enough that they are what Lechterman calls "constitutive rules about what 'counts' as engaging in the practice" (unpublished, 5): e.g. if someone deliberately prints a falsehood, then whatever they are doing, it is not *real journalism*.

[7] This requirement also provides us with the resources for excluding investigation of practices that are widely viewed as immoral, such as being an assassin or pickpocket, since their effects cannot be justified to those they affect. James says as much, endorsing a contractualist account of justification, rejecting practices that involve "dominance, negligence, or exclusion" (2005: 310).

identify the goals of their domain.[8] Ideally, these prompts would converge on one or a set of general answers:

- Why do people choose to go into this field over others?[9]
- What are people within this domain hoping to contribute to society?
- How do the people working in this domain praise themselves, e.g., in their advertisements, award ceremonies, or public statements?
- What benefits do consumers, users, or broader society expect these domains to furnish?
- What is the point of these domains in the eyes of outsiders?

Finally, we can reason from the goals of a domain down to the norms that govern the appropriateness of behavior within the domain. We call these the *values* of a domain and define a value as "an aspect of our activity within a domain that we wish to promote or preserve; features or qualities of our actions that merit attention while we are pursuing our goals." This is broadly consonant with other discussions of values in the technology ethics literature (see Van de Poel 2013, especially pgs. 262 and following; and the other authors cited there, e.g., Anderson 1993 and Dancy 2005). For example, a teacher might have the *goal* of spurring her students' interest in her field, but she might *value* honesty in doing so. Valuing honesty means that certain ways of spurring her students' interest, e.g., lying, misleading, or acting in bad faith, are unacceptable. Values can be thought of as providing constraints that rule out certain methods of accomplishing our goals, or reasons that count in favor of certain methods over others.

Attending to both goals and values is a way to more exhaustively investigate the ethical dimensions of an ML system, surfacing factors we might overlook if all we focus on are the goals at hand. The goals associated with a particular application may be served by it while violating the values of the domain. For example, a system that predicts outcomes of court cases might do a great job for an individual attorney but, in doing so, may violate the important value of fairness in the criminal justice system (e.g., Shaikh et al. 2020).

This stage of the process often benefits from consulting the documents promulgated by a domain's professional bodies, which explicitly lay out a profession's aspirations and values.[10] Still, *professions* are narrower than *occupations*, and still more narrow than *domains* as we understand them. But for those domains containing mature professions, this task is easiest. We have the advantage in this project of investigating domains, almost exclusively, that have regulative professional bodies, such as law (the American Bar Association), journalism (the Society of Professional Journalists), and medicine (the American Medical Association).

Before we move on, we will address a cluster of concerns about the theoretical viability of attributing goals and values to domains. As one participant at our 2020 workshop put it: "*People* have goals; *domains* do not have goals." Second, we might worry that people *within* a domain have different goals. Third, individuals within a domain might have different goals than the goals we attribute to the domain itself. The people in a company who answer phones or process invoices might not have any particularly lofty goals at all. All these objections express skepticism that we can treat domains as if they were *monolithic agents with univocal intentions* when, of course, they are not.

Nonetheless, we are confident that we can identify the benefits that domains purport to provide. The more seasoned practitioners in a domain should be able, upon reflection, to provide at least some explanation of and justification for what they are doing with their lives. Still, consensus would be a quixotic goal and we should be satisfied if we can arrive at answers that *most reasonable practitioners* could accept and that enjoy *wide endorsement* among the domain's practitioners. Beitz is especially helpful here: what we seek is "a facially reasonable conception of the practice's aim [and values] formulated so as to make sense of as many of the central normative elements as possible within the familiar interpretive constraints of consistency, coherence, and simplicity" (2009: 108). We are buoyed by the success of this method in some domains, for example, in the history of the professionalization of journalism, and the coalescence of journalists worldwide around a broadly shared understanding of the goals and values of their work.[11]

[8] When a domain has multiple goals, we must also consider the possibility that implementing a model to optimize for one goal could undermine the peripheral goals of the domain (Mesthene 1997), for example, by efficiently selecting applicants for higher education but reducing the diversity of the students selected.

[9] While individuating domains has always posed a challenge, the most promising method for doing so remains appealing to their respective goals. Activities with goals that cannot be reconciled—marketing and journalism, for example—ought to be kept distinct.

[10] In the sociology of professions, the authors often take the existence of a code of ethics, which articulates shared understandings and expectations of appropriate behavior, to be crucial for professionalization. See Wilensky (1964) for a classic treatment, and Abbott (1991) and Hall (1968) for other classic discussions of the 'process' model of professionalization. See Forsyth and Danisiewicz (1985) for a literature review and defense of alternative theories of professionalization.

[11] See Deuze (2005), which traces the history of journalism's self-perception and the formation of its professional identity, which is "kept together by the social cement of an occupational ideology" (2005: 442). See also Weaver (1998: 456), who argues that the late-twentieth century "consolidation" of journalistic values even stretched across national borders.

3.5.2 Applying goals and values to case studies

It is helpful to approach the language of goals and values by examining several contentious examples of the use of machine learning. For the first example, consider the case of COMPAS. In 2016, ProPublica published a bombshell investigative story which galvanized the public conversation around the fairness, accuracy, and transparency of algorithms (Angwin et al. 2016). ProPublica revealed that a company in Florida, NorthPointe, developed an algorithm called COMPAS for assessing the risk that someone convicted of a crime would reoffend, based on 100 + factors. This algorithm was used during parole hearings to help parole boards decide whether to release a convict eligible for parole. ProPublica showed that this model was biased, tending to *overestimate* the risk of black convicts reoffending and *underestimate* the risk of white convicts reoffending. Thus, the system seemed biased in precisely the way that the criminal justice system has been historically biased against people of color. If ProPublica's analysis is correct—which is controversial (Corbett-Davies et al. 2019)—this algorithm is clearly problematic because it is *unfair*.

However, imagine that the predictions that COMPAS yielded were *perfect*, i.e., that it could perfectly predict whether someone who is up for parole would commit a crime if they were released from jail. Does this algorithm still seem problematic? The answer still seems to be, Yes: most people would still have some anxiety or unease about using this algorithm. Why is this? Perhaps because what COMPAS is doing is *inappropriate given the goals and values of the domain within which it is deployed*. The goal of the parole system is not *primarily* to predict whether someone is going to commit a crime.[12] The purpose of parole is usually thought to be one of rehabilitating and reintegrating former prisoners (Lynch 2000; Simon 1993), to release them for a period of "supervised readjustment" (Wilcox 1929: 346). However, the COMPAS model *naturally invites* the members of the parole board to think about whether someone is likely to commit a crime if they were released. Thus, the concern was that the model served to redirect the institution away from the goals of its domain.

Consider next a comparison between two uses of recommendation systems. These systems recommend things to the user that they would enjoy or that seem relevant to them. Users encounter recommendation systems anytime they log onto Netflix, Amazon, Spotify, etc. Users also encounter these systems on Facebook when they are shown ads that Facebook's advertising algorithms recommend. This seems appropriate given the goal of advertising: presumably something like *telling consumers about products that will improve their lives*. (Note that, in this case, the language of goals and values is useful to articulate a *defense* of a use of machine learning.)

However, that use of machine learning becomes problematic when the same algorithm is used to curate the *information about the external world* users see on their Facebook newsfeed. This seems problematic because there is a significant difference between the goal of advertising and the goal of journalism.[13] The purpose of journalism is *at least in part* to tell us things that we *need* to know, even if we do not *want* to know them or *do not know* that we need to know them. If users are only shown things about the world that they *want* to see, that begins to undermine the ability of journalism to deliver its characteristic value to society. This shows that we can distinguish between uses of machine learning that are appropriate or inappropriate by examining the goals of the domain in which they are being deployed—even if the same basic system is used in each case. The goals of advertising are served by systems that give people what they want and establish brand loyalty. As Facebook's own research has shown, the same algorithms lead to information that aligns with users' beliefs (what they "want" to read) and polarization of thought (or loyalty to ideas) (Bidar 2021).

4 Major findings

Two virtual workshops were held, in November 2020 and June 2021, to further develop and refine the evaluation framework, piloting the framework in several domains. The goal was to assess whether this core structure could be applicable and useful across these domains and various use cases and to identify concrete projects to continue refining the framework. Below we report primarily on the findings from November 2020.

[12] One reviewer for this journal has suggested that we consider the "intermediate goals" of the parole system. We take up—but ultimately reject—the suggestion of incorporating "intermediate goals" into our framework below when discussing profit as a candidate intermediate goal.

[13] A reviewer for this journal raised an incisive objection at this point, asking why we should consider social media companies to be participants in the domain of journalism at all. Could they, for example, argue that they are not part of this domain, and thus not bound by its traditional goals and values? This raises particularly thorny questions about drawing the boundaries of a domain, and we admit this requires more thought. However, in this case, we think this attribution is fair for two reasons. First, Pew found in 2021 that Facebook leads social media sites as a source of news for users: around half of its users (and 31% of Americans) "regularly get news" about current events from Facebook; therefore, Facebook is furnishing the same good that journalism traditionally has for society (Pew Research Center 2021: 4–5). Second, by Mark Zuckerberg's own admission, Facebook aspires to furnish users' "primary news experience" (Owen 2015).

Six core domains were represented at the workshop with breakout groups of subject matter experts and practitioners: Law, Journalism, Medicine, Business, Manufacturing and Robotics. Medicine was further subdivided into a group focused on Imaging, which is further advanced in the utilization of ML methods, and a group focused on Text, Sensors and Omics, which are emerging areas of ML utilization in Medicine. Researchers and participants were invited to have a mix of expertise in each domain focus group, with individuals who focus on fundamental machine-learning model development, those who work on creating specific applications of ML models in practice, those who are practitioners in the domain, and others who focus on the study of impact of AI systems. It was intended that this mix of expertise and perspective would create opportunities for idea exchange and constructive, challenging dialogue. Fifty-two researchers and practitioners participated in the workshop, with each breakout group ranging from five to eight individuals.

The workshop consisted of a set of short focused presentations on each major framework component as well as a keynote presentation on the overarching positive and negative impact potential of machine learning. During breakout sessions, each group was provided with a set of two or three case study examples to examine in light of the evaluation framework.

4.1 Goals and values identified

The following table (Table 1) reports the goals and values that were identified during our November 2020 workshop.

4.2 Negotiating trade-offs

A crucial aspect of evaluating the human impacts of a use of machine learning is appreciating that goals and values might conflict. When this happens, it is necessary to resolve those conflicts, which can sometimes involve trading off one goal or value against another (Van de Poel 2015; Van de Poel and Royakkers 2011). This tension surfaced as a major hurdle during our workshops.

What makes some design *good* is often a matter of its performance along several dimensions. Engineers are familiar with the trade-offs that are sometimes required between, for example, safety and esthetics, or environmental sustainability and power. The safest car might not be the prettiest; and the most environmentally sustainable car is likely to be made out of materials that are not the strongest known. Analogous choices confront computer scientists working in machine learning. Machine learning requires us to optimize for particular goals, and often requires us to make trade-offs. For example, a well-recognized trade-off in machine learning is between accuracy and fairness (where fairness is defined mathematically, without any moral implications intended). Another is between power and transparency: more powerful machine-learning models tend to be more opaque.

Among those who work in technology ethics, there is a clear consensus that negotiating these trade-offs is an outstanding challenge. This is attested to by the rise in approaches and subdisciplines such as "value sensitive design" (Friedman 1996; Friedman et al. 2002), "responsible innovation" (Owen et al. 2013; Van den Hoven et al. 2014), "ethical technology assessment" (eTA) (Palm and Hansson 2006; Kiran et al. 2015), and others.

The framework we propose applies a moral lens to this process of navigating trade-offs. Part of the conceptual function of identifying values is to help us articulate the necessary trade-offs in a moralized language and to evaluate whether specific applications of machine learning are having an acceptable impact on humanity. The goal should be to settle on one solution that belongs to a larger *set* of acceptable solutions, and to be able to defend that choice in moral language, being cognizant of the trade-offs that are required.

But this raises vexing issues about just *how* we can strike a balance between multiple values that we care about. When optimizing along multiple dimensions, it can be difficult or impossible to know which solution is 'best.' These difficulties are intensified when we move from an empirical field like computer science into one that is non-empirical, like moral philosophy. In a word, work remains to be done to *operationalize* the negotiation of these trade-offs.[14] Addressing this challenge is, obviously, outside of the scope of this paper, although promising work has been done on this question recently. See van de Poel (2015: 90) and especially Van de Kaa (2020: 477) for several such methods for resolving trade-offs.

4.3 Confronting the conflict between ethics and profit

Perhaps the primary concern voiced by the participants was that our conception of goals and values as *aspirational* ignores an important motivation that many people have: the pursuit of money. As a matter of descriptive fact, this is often

[14] See, for example, one of the "top takeaways" of the 2021 Artificial Intelligence Index Report from Stanford University's Human-Centered AI Institute: "Though a number of groups are producing a range of qualitative or normative outputs in the AI ethics domain, the field generally lacks benchmarks that can be used to measure or assess the relationship between broader societal discussions about technology development and the development of the technology itself" (4). On page 127, the report elaborates, "Figuring out how to create more quantitative data presents a challenge for the research community, but it is a useful one to focus on. Policymakers are keenly aware of ethical concerns pertaining to AI, but it is easier for them to manage what they can measure, so finding ways to translate qualitative arguments into quantitative data is an essential step in the process.".

Table 1 Goals and values identified during our November 2020 half-day workshop

Domain	Goals	Values
Education	· Assessing students' potential · Encouraging students to do their best · Developing human capital · Enabling social mobility	· Allowing students to participate in their own success · Diversity · Fairness, meritocracy, and grit · Rewarding focus and preparation · Equal playing field · Giving students second chances
Journalism	· Contributing to the public's knowledge and understanding · Serving as a check on those in power	· Fairness · Balance in coverage
Law	· Equal justice under the law for everyone · To correct inequities · Retribution · Securing compensation for the wronged · Deterring and discouraging crime · Maintain order · Discovering the truth	· Everyone gets a fair shot · Non-arbitrariness: principled consistency · Equality · Accountability · Transparency · Efficiency · Due process · Neutrality and objectivity
Manufacturing	· Help clients deliver products to their customers	· Safety · Quality · Service · Cost · Employee satisfaction
Medicine	· Diagnosing patients and helping to make them better · Predicting clinical outcomes · Scientific discovery	· Accuracy · Affordability · Speed · Fairness · Beneficence · Non-maleficence · Scientific rigor
Robotics[a]	· Training machines to move through the world as embodied intelligences · Creating machines to do useful things in the world · Exploring and expanding what is possible	· Safety · Utility

[a]Robotics is an interesting and, perhaps, peculiar case. The discipline of robotics creates tools that cut across multiple other domains. This suggests that robotics could, to an extent, inherit the goals and values of the domain into which it is deployed: the goals of medical robotics, for example, might differ from the goals of agricultural robotics or autonomous vehicles, etc.

a goal of the domains we discussed, and a primary motivation of implementing machine-learning systems.

This is an old problem, surfacing yet again in a new guise: What happens when doing the right thing is costly? We do not expect to be able to resolve this tension as part of this project. However, a satisfactory framework for assessing the human impact of machine learning must provide some guidance for navigating this tension and situating the role of this framework within the more general task of businesses to decide how they balance competing concerns.

There are several ways of responding. First, we should acknowledge that financial considerations constrain the way many companies pursue their goals. For that reason, we could simply include profit as a *value* of these domains, since values are always to be weighed against one another during decision making. Other considerations would thereby be prevented from overriding the pursuit of profit entirely.

But there are several other reasons to exclude financial considerations from this framework altogether. First, recall that above we suggested that when individuating domains, we should consider the *characteristic benefit* that these domains provide. How do we distinguish, for example, between literature and journalism? One of these is supposed to deliver the benefit of helpful information about current

events, i.e., to function as "the first draft of history." This distinguishes journalism from literature. According to this concept of a domain, then, money cannot be the goal of a domain because it is not *unique* to any one domain (Plato and Reeve 2004: 345e–346d).

Second, our framework is meant to serve, in part, as a counterweight to a single-minded focus on profit, which can be seen as a source of problematic design decisions in many of the examples discussed above.

Finally, the paradigmatic goals of the domains discussed here tend to be *intrinsic goods*, that is, things that are desirable for their own sake. The domain of medicine provides patients with health; the domain of journalism provides readers with truth or understanding about the world. These examples have been seriously entertained as intrinsically valuable. Money is not this way: it is an instrumental good, or something that we only pursue because it helps us attain other things that we want, such as health or knowledge. Thus, profit is not properly conceived as a goal of a domain like the other goals under consideration here.[15,16]

Oftentimes, "good ethics is good business," and there at least *tends* to be an alignment between profit and ethics. But there is no guarantee that this is true and there are well-known structural aspects of capitalism that allow companies to shift the harmful consequences of bad decisions onto society. The negative human impacts of artificial intelligence can be seen as one more kind of *negative externality*. In the absence of regulation to mitigate these externalities, the responsibility for minimizing them "follows the money." The design of many machine-learning systems tends to cater to those who are paying for the design and, as a result, can be insufficiently sensitive to the interests of those who are *subjected to* them. Those who are responsible for the design and deployment of machine-learning systems should understand how to anticipate the negative impacts of their designs, and our framework is precisely an effort to furnish technologists with such a tool. It is our view that those who design and implement machine-learning systems are *no longer allowed to plead ignorance* in the face of the profound negative impacts of their designs.

5 The way forward: next steps

5.1 Fine-tuning domains, goals and values

Move towards consensus goals and values for the domains under consideration. In our brief workshops, we were not realistically able to arrive at such a consensus. Instead, we demonstrated a proof of concept that, given such goals and values, we could make meaningful criticisms of particular machine-learning systems.

Consider framing goals and values in the language of constraints, requirements, or other concepts imported from engineering. Some participants suggested that it would be valuable to add a category of "constraints" in addition to goals and values. Including explicit constraints in the framework could help narrow the set of alternatives under consideration. Additionally, this might be helpful for accommodating commercial concerns in the design and development of machine learning.

Expand the list of domains under investigation. Expanding the list of domains—to include, for example, the military, education, finance, and policing—might illuminate unexpected connections between domains, such as commonalities in goals or values that can be used to develop a "unified theory" of AI ethics.

Identify possible points of friction between goals, values, individuals, and society. Some participants raised questions about the goals and values that society might have, on the one hand, and that individuals operating *within* domains might have, on the other hand. The choice of *domains* as the unit of analysis could overlook the nature of domains as dynamic collections of individuals, each with their own desires, habits, and so on—though this may be a necessary theoretical cost.

Acknowledge the cultural differences in goals and values. Shepherding a promising future for machine learning, in which people around the world are able to share in its benefits, requires acknowledging deep-seated differences or outright disagreements about the goals and values of a domain. Future iterations of this work should explicitly state when goals and values may be culturally bound or, on the other hand, when they range across cultures.

5.2 Specifying trade-offs

A high priority ought to be investigating the extent to which existing methods in engineering for negotiating trade-offs

[15] One possible exception to this is that profit might be the intrinsic good of some domains like finance or stock trading. Though, even here, the goal is to maximize return for one's clients, which is distinct from an institution pursuing money for its own sake.

[16] One reviewer has suggested that we include profitability as an "intermediate goal" of a domain, which is sought in pursuit of a further goal. We appreciate this suggestion but disagree. First, we believe that we can accommodate a concern for profitability by considering profit to be a value of a domain, rather than a goal. Recall that the role of values in the framework is analogous to a constraint: applications of ML which are not profitable are not acceptable, in the same way that other applications that violate a value of the relevant domain are unacceptable. Second, given this, we believe that inserting an additional layer like this to the evaluative component of the framework would add complexity beyond what is necessary to account for the concern of intermediate goals.

can be ported over to help negotiate *ethical* trade-offs in design.

First, an ideal method for negotiating trade-offs would not require expertise in mathematics or computer science. The most promising ways of arriving at broad-based agreement about trade-offs between values will require acknowledging input from stakeholders across communities, affected populations, disciplines, etc. It would be unwise to expect the participants in these conversations to have deep expertise in mathematics or computer science. Second, an ideal method would be context-sensitive. It should be responsive to the fact that the relevant values or their weights might differ between professional, cultural, or national contexts. It should, therefore, be flexible to accommodate different constraints, rankings, and trade-offs.

5.3 Translating goals and values into measurable human impact

Continue to translate values and goals into concrete, actionable, quantifiable impacts. Operationalizing any framework to evaluate the human impact of machine learning will require that we can assign precise values to the moral issues under consideration. This will help with the integration of these goals and values into machine-learning projects and into decision-making processes. It will also help when it comes time to fine-tune the weights of these goals and values, for example, between companies within a domain. This is a vexing project. Some goals and values lend themselves to quantification relatively easily—such as estimating the number of stakeholders affected by a decision. Others less so—for example, quantifying the 'badness' of inequality.

Further explore higher-level considerations which arch over the use of machine learning in different domains, and which can guide the implementation of machine learning and integration into existing processes. This will involve considering several aspects of a domain and of particular algorithms. For example, calculating the cost of false positives and negatives reveals that different domains have different risk tolerances: it is much worse for an algorithm to deliver a false positive if it is making recidivism predictions for a parole board than if the algorithm is delivering targeted ads.

Develop guidelines for the integration of machine learning into concrete contexts. Much of the human impact of machine learning turns on what happens before and after a model is developed, trained and deployed. Specifically, much depends on the reliability and equity of the data that are selected before the model is trained; and much depends on which human decisions the model's verdicts are driving. Any evaluation of the human impact of machine learning would be lacking if it neglected the importance of deploying machine-learning models with a constant eye towards the accessibility, accountability, transparency, and equity of the domains within which they live and operate.

Perhaps the most important outcome of our workshops was a qualified optimism about this general approach, which still strikes us as promising, even if there is additional conceptual and procedural work to be done. We are hopeful that a future version of this framework, which is powerful, contextually sensitive, and action-guiding, will make an important contribution to the ongoing development of methods to anticipate the human impacts of machine learning.

Acknowledgements The funding for this research was provided by Underwriters Laboratories Inc. through the Center for Advancing Safety of Machine Intelligence, as well as through National Science Foundation award 1917707, "Artificial Intelligence and Predictive Policing: An Ethical Analysis." We owe a great thanks to the participants at our 2020 and 2021 workshops for their stimulating and enriching discussion. In particular, we would like to thank the presenters at those workshops whose contributions strengthened and shaped the framework we articulate here: Mary "Missy" Cummings, David Danks, Nicholas Diakopoulos, Jim Guszcza, Brent Hecht, Abigail Jacobs, Patrick Lin, and Julia Stoyanovich. Two anonymous reviewers for this journal pressed us to clarify crucial points and have strengthened the paper. Finally, David Askay and Ted Lechterman provided indispensable guidance for improving our conception of social domains.

References

Abbott A (1991) The order of professionalization: an empirical analysis. Work Occup 18(4):355–384

Alikhademi K et al (2021) A review of predictive policing from the perspective of fairness. Artif Intell Law 30:1–17

Allen C, Smit I, Wallach W (2005) Artificial morality: top-down, bottom-up, and hybrid approaches. Ethics Inf Technol 7(3):149–155

Anderson E (1993) Value in ethics and economics. Harvard University Press, Cambridge

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine Bias. ProPublica, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 28 May 2021

Barocas S, Selbst AD (2016) Big data's disparate impact. Calif L Rev 104:671

BBC News (2018) Amazon Scrapped 'sexist AI' Tool. *BBC News*. https://www.bbc.com/news/technology-45809919. Accessed 12 Nov 2021

Beitz CR (2009) The idea of human rights. Oxford University Press, Oxford

Bidar M (2021) Liberals to 'Moscow Mitch,' Conservatives to QAnon: Facebook Researchers Saw How Its Algorithms Led to Misinformation." CBS News. https://www.cbsnews.com/news/facebook-algorithm-news-feed-conservatives-liberals-india/
Brey P (2004) Ethical aspects of facial recognition systems in public places. J Inf Commun Ethics Soc. https://doi.org/10.1108/14779960480000246
Canca C (2020) Operationalizing AI ethics principles. Commun ACM 63(12):18–21. https://doi.org/10.1145/3430368
Corbett-Davies S et al (2019) A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It's Actually Not That Clear. Washington Post. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/. Accessed 28 May 2021
Crespo R (2016) Aristotle on agency, habits and institutions. J Inst Econ 12(4):867–884
Dancy J (2005) Should we pass the buck? In: Rønnow-Rasmussen T, Zimmerman MJ (eds) Recent work on intrinsic value. Springer, Dordrecht, pp 33–44
Deuze M (2005) What is journalism?: Professional identity and ideology of journalists reconsidered. Journalism 6(4):442–464
Dworkin R (1986) Law's empire. Belknap Press of Harvard Univ. Press, Cambridge
Erman E, Möller N (2015) Practices and principles: on the methodological turn in political theory. Philos Compass 10(8):533–546
Erman E, Möller N (2016) What distinguishes the practice-dependent approach to justice? Philos Soc Crit 42(1):3–23
Fjeld J, Nele A, Hannah H, Adam N, Madhulika S (2020) Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. Berkman Klein Center Research Publication No. 2020-1, Available at SSRN: https://ssrn.com/abstract=3518482 or https://doi.org/10.2139/ssrn.3518482
Forsyth PB, Danisiewicz TJ (1985) Toward a theory of professionalization. Work Occup 12(1):59–76
Friedman B (1996) Value-sensitive design. Interactions 3(6):16–23
Friedman B, Kahn P, Borning A (2002) Value sensitive design: theory and methods. University of Washington technical report
Goodman B, Flaxman S (2017) European Union regulations on algorithmic decision-making and a "right to explanation." AI Mag 38(3):50–57
Hall RH (1968) Professionalization and bureaucratization. Am Sociol Rev 33(1):92–104. https://doi.org/10.2307/2092242
Hindriks F, Guala F (2015) Institutions, rules, and equilibria: a unified theory. J Inst Econ 11(3):459–480
Hodgson GM (2006) What are institutions? J Econ Issues 40(1):1–25
Hodgson GM (2015) On defining institutions: rules versus equilibria. J Inst Econ 11(3):497–505
IEEE (2018) Ethically Aligned Design." Version 2: For public comment. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Available at https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf. Accessed 27 May 2021
James A (2005) Constructing justice for existing practice: Rawls and the status quo. Philos Public Aff 33(3):281–316
Jubb R (2016) 'Recover it from the facts as we know them': practice-dependence's predecessors. J Moral Philos 13(1):77–99
Kaminski ME (2019) The right to explanation, explained. Berkeley Tech LJ 34:189
Kiran AH, Oudshoorn N, Verbeek P-P (2015) Beyond checklists: toward an ethical-constructive technology assessment. J Responsible Innov 2(1):5–19
Kirkpatrick J, Hahn EN, Haufler AJ (2017) Trust and Human-Robot Interactions. In: Lin P, Jenkins R, Abney K (eds) Robot Ethics 2.0: from Autonomous Cars to Artificial Intelligence
Kroes P et al (2006) Treating socio-technical systems as engineering systems: some conceptual problems. Syst Res Behav Sci 23(6):803–814
Lamarque P (2010) Wittgenstein, literature, and the idea of a practice. Br J Aesthet 50(4):375–388
Lynch M (2000) Rehabilitation as rhetoric: the ideal of reformation in contemporary parole discourse and practices. Punishment Soc 2(1):40–65
MacIntyre A (1981) After virtue: a study in moral theory. University of Notre Dame Press, Notre Dame
MacIntyre AC (1988) Whose justice? Which rationality? University of Notre Dame Press, Notre Dame
Madaio MA, Stark L, Vaughan JW, Wallach H (2020) Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–14. Honolulu HI USA: ACM https://doi.org/10.1145/3313831.3376445
Max R, Kriebitz A, Von Websky C (2020) Ethical considerations about the implications of artificial intelligence in finance. In: Handbook on Ethics in Finance. pp 1–16
Mesthene EG (1997) The role of technology in society. In: Schrder-Frechette K (ed) Technology and values. pp 71–85
Mittelstadt B (2019) AI Ethics–Too Principled to Fail? CoRR arXiv:1906.06668 (2019)
Nissenbaum H (2004) Privacy as contextual integrity. Wash Law Rev 79:119
Nissenbaum H (2011) A contextual approach to privacy online. Daedalus 140(4):32–48
Owen LH (2015) Mark Zuckerberg Has Thoughts on the Future of News on Facebook. Nieman Lab. https://www.niemanlab.org/2015/06/mark-zuckerberg-has-thoughts-on-the-future-of-news-on-facebook/. Accessed 13 Nov 2021
Owen R et al (2013) A framework for responsible innovation. In: Responsible innovation: managing the responsible emergence of science and innovation in society, vol 31, pp 27–50
Palm E, Hansson SO (2006) The case for ethical technology assessment (eTA). Technol Forecast Soc Chang 73(5):543–558
Pew Research Center (2021) News Consumption Across Social Media in 2021.https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/. Accessed 13 Nov 2021
Plato, Reeve CDC (2004) Republic. Indianapolis: Hackett Pub. Co
Rawls J (1955) Two concepts of rules. Philos Rev 64(1):3–32
Rhodes RAW et al (eds) (2006) The oxford handbook of political institutions. Oxford University Press, Oxford
Richardson HS (1997) Practical reasoning about final ends. Cambridge University Press, Cambridge
Rodgers S (2021) Themed issue introduction: promises and perils of artificial intelligence and advertising. J Advert 50(1):1–10. https://doi.org/10.1080/00913367.2020.1868233
Searle J (2005) What is an Institution? J Inst Econ 1(1):1–22
Selbst A, Powles J (2018) Meaningful Information and the Right to Explanation. In: Conference on fairness, accountability and transparency. PMLR
Selinger E, Leong B (2021) The ethics of facial recognition technology. In: Véliz C (ed) Forthcoming in The Oxford Handbook of Digital Ethics
Shaikh RA, Sahu TP, Anand V (2020) Predicting outcomes of legal cases based on legal factors using classifiers. Proc Comput Sci 167:2393–2402. https://doi.org/10.1016/j.procs.2020.03.292
Simon J (1993) Poor discipline. University of Chicago Press, Chicago
Tatum JS (1997) the political construction of technology: a call for constructive technology assessment. In: Shrader-Frechette K (ed) Technology and values. pp 115

Van de Poel I (2013) Translating values into design requirements. In: Philosophy and engineering: reflections on practice, principles and process. Springer, Dordrecht, pp 253–266
Van de Poel I (2015) Conflicting values in design for values. In: Handbook of ethics, values, and technological design: sources, theory, values and application domains, pp 89–116
Van de Poel I (2020) Embedding values in artificial intelligence (AI) systems. Minds Mach 30(3):385–409
Van de Poel I, Royakkers L (2011) Ethics, technology, and engineering: an introduction. Wiley, Hoboken
Van den Hoven J et al (2014) Responsible innovation. In: Third international conference on responsible innovation, vol 22
Van de Kaa G et al (2020) How to weigh values in value sensitive design: A best worst method approach for the case of smart metering. Sci Eng Ethics 26(1):475–494
Verbeek P-P (2005) What things do: philosophical reflections on technology, agency, and design. Pennsylvania State Univ. Press, University Park
Walzer M (2008) Spheres of justice: a defense of pluralism and equality. Basic books
Weaver DH (ed) (1998) The global journalist: news people around the world. Hampton Press, New Jersey
Wieringa M (2020) What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In: Proceedings of the 2020 conference on fairness, accountability, and transparency
Wilcox C (1929) Parole: principles and practice. Am Inst Crim L Criminol 20:345
Wilensky HL (1964) The Professionalization of Everyone? Am J Sociol 70(2):137–158
Zhang D, Mishra S, Brynjolfsson E, Etchemendy J, Ganguli D, Grosz B, Lyons T, Manyika J, Niebles JC, Sellitto M, Shoham Y, Clark J, Perrault R (2021) The AI Index 2021 Annual Report. AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford