Northwestern | MACHINE LEARNING IMPACT INITIATIVE

A Framework for the Design and Evaluation of Machine Learning Applications

Northwestern University Machine Learning Impact Initiative

September 2021

This document was compiled by Kristian J. Hammond, Ryan Jenkins, Leilani H. Gilpin, and Sarah Loehr with assistance from Mohammed A. Alam, Alexander Einarsson, Andong L. Li Zhao, Andrew R. Paley, and Marko Sterbentz. The content reflects materials and meetings that were held as part of the Machine Learning Impact Initiative in 2020 and 2021, with the participation of a network of researchers and practitioners. See the Machine Learning Impact Initiative Summary Report for a full list of all who participated and engaged in these processes.

This work was supported through a gift from Underwriters Laboratories.

Table of Contents

Introduction

Machine Learning – the ability of computers to process and learn from masses of data – is being utilized in applications that impact nearly every aspect of human life and society. Unfortunately, the speed of deployment of these applications has outpaced our ability to evaluate their impact, both positive and negative. Even more troublesome, while there is strong agreement that evaluation, testing, and possible regulation of these technologies is needed, there has been little work aimed at operationalizing these steps. As a result, we are faced with an ever-growing set of applications that are having tremendous impact on society and human health and safety, but few practical mechanisms for understanding either that impact or its sources.

As a first step in addressing this problem, the Northwestern/UL Machine Learning Impact Initiative (MLII) worked to develop an *Evaluation Framework* that can be used to operationalize the evaluation and testing of the impact of existing applications and guide the design and development of new ones.

The Framework is the result of an unpacking of the processes and materials that are used to develop machine learning systems so each element can be examined independently. This decomposition parallels the process of developing applications based on ML in general. The decoupling allows us to build a standardized and operationalizable model for examining ML applications from the perspective of their individual components and their impact, a step that dramatically reduces the complexity of the problem. This decoupling also clarifies the nature of the problems that can arise in ML development at the component level and how to identify them.

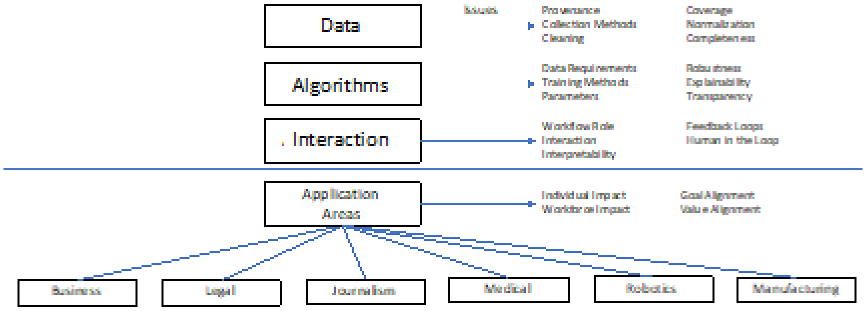
The Framework divides the evaluation process into two primary components: gathering the facts pertaining to the system being assessed and evaluating those against domain-level goals and values. This division supports a two-phased model. First there is an initial judgement-free fact gathering process aimed at articulating the core features of a system. This information-gathering phase is followed by a process that evaluates the impact that the specifics of those facts have in relation to the goals and values of the domain and the human outcomes that they determine.

Facts: In looking at the impact of ML systems, we need to consider three primary components: the data and how it was sourced and manipulated, the central ML algorithms and how they were applied, and the ways in which the resulting systems are designed to interact with human users. While each of these components feeds into the others, each has different concerns and issues that must be considered.

Evaluation: Given a collection of facts drawn from these three areas, the domain in which an application will be deployed defines a core set of explicit goals and implicit values that must be considered. As different domains have different goals and values, a single application that might be acceptable in one domain might very well be ruled out of others.

This separation of concerns clarifies the overall process and allows the Framework to be applied across any application and domain.

This approach is based on a simple idea: the features of a system are different from the evaluation of those features. That a data set does not reflect the distribution of examples in the real world is a fact. That this distribution results in a bias that might cause harm in a legal setting, is the evaluation. Before we can focus on the evaluation, we must first establish the core facts.



Decoupling Facts from Evaluation

The primary effect of this division is to separate those elements upon which there can be objective agreement (i.e., the facts of the matter) from those for which there may be some discussion (i.e., the evaluation). This both clarifies and simplifies the process by focusing the discussion on the impact of a set of features rather than the existence of the features themselves.

Access to facts associated with a particular application may be blocked for a variety of reasons. Algorithms and data may be proprietary. The details of data gathering and normalization or the process behind the selection of features within that data may be forgotten. Decisions about training of a system may have been *ad hoc* to begin with and undocumented.

No matter the cause, the functional pressure of the Framework provides a focus that allows us to both identify what information is needed and to uncover those situations where the unknowns themselves give rise to potential research issues. The Framework defines the questions that need to be asked. In turn, those questions that cannot be answered with current methodologies define the research needed to answer them.

To illustrate this Framework process, consider an example of the problem that arises in training sets with the use of proxy measurements. In many cases, objectives of a process or system may be difficult to measure directly; however, there may be other proxy metrics and data that are reasonably easy to attain and to utilize at the volume of data that is necessary for machine learning systems. Employee performance might be approximated using sales figures. Likewise, money spent on healthcare might be used as a proxy for health of a patient or the severity of a medical condition (Obermeyer et al. 2019). In these cases, the available proxy data might be used to train a machine learning system. But, if that proxy is not aligned with the features that it is intended to represent, the resulting system might be skewed, in these cases, biased against historically underserved groups, and as a result may exacerbate harms to individuals in those groups in the future. It is common today for applications like this to have promising results in testing, to be deployed to production use and for these issues to be discovered only with post-hoc analysis. The framework could allow us to ask the questions necessary to identify the gaps in research and the techniques that need to be developed in order to avoid, prevent, mitigate and evaluate the issues regarding the viability of using a proxy.

Motivation and Context

Efforts to guide the ethical development of artificial intelligence have proliferated at an astonishing pace over the last few years — various organizations have published at least 96 reports, guidelines, sets of best practices, and so on since 2018 alone, and another 21 before that (Zhang et al. 2021: 130). While these reports have been produced by governments, professional organizations, NGOs, private corporations, universities, think tanks, and others, they are largely top-down (Allen, Smit, and Wallach 2005): they put forth general principles at the highest level which can arch over the development and deployment of artificial intelligence in any domain. Common themes among these guidelines are fairness, accountability, transparency, and explainability (the so-called 'FATE' concepts); 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 General Data Protection Regulation (GDPR), for example, which guarantees a citizen's right to explanation (Goodman and Flaxman 2017), there has been a flurry of activity exploring the nature of explanation and the nature of this purported duty held by governments, corporations, or others towards data subjects (Kaminski 2019; Selbst and Powles 2018).

As the development and refinement of AI techniques continues apace, identifying these overarching values and investigating their nature is clearly important. But it comes with several costs which are also becoming clearer. First, what these frameworks boast in generality they sacrifice in power and capability for action guidance. We are not the only ones to share this view. See, for example, Zhang et al., who bemoan that "the vague and abstract nature of those principles fails to offer direction on how to implement AI-related ethics guidelines" (2021: 129). For one thing, these principles require a tremendous amount of work to operationalize, and they have led to disagreements even 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).

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 additional obstacle to their operationalization. There are significant and reasonable disagreements between domains about the nature and relevance 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. inscrutable to human users), 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 artificial intelligence: e.g. advertising recommendation systems online (Rodgers 2021); facial recognition in law enforcement (Brey 2004; Raji et al. 2020; Selinger and Leong 2021); prediction models in finance (Max, Kriebitz, and Von Websky 2020; Davis, Kumiega, and Van Vliet 2013), and so on. 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. It is difficult to generalize these findings for AI broadly, and often for other applications in the same domain. Moreover, it would be onerous to examine the impacts of every single instance of use of machine learning in society.

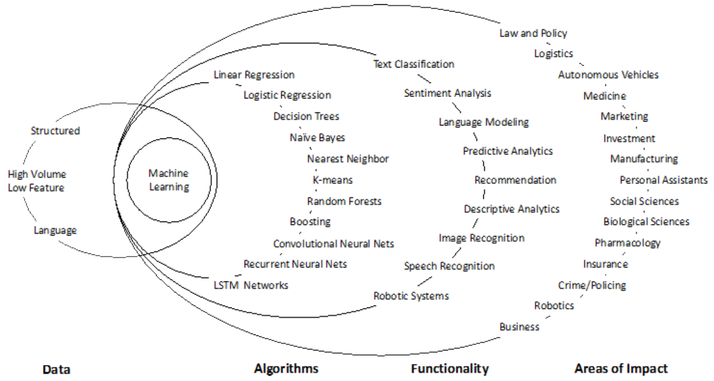
Spurred by these observations, we are attempting to thread the needle by developing what we characterize as a middle-out approach. This approach takes social domains as the appropriate level of analysis, identifies the individual goals and values of those domains, and then explores how particular implementations 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 institutions 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 a function of the relevant goals and values and its embedded use.

Framework Structure

As described above, this Framework for the evaluation of the impact of systems based on machine learning divides the task into two phases: extraction of the facts related to the design and development of systems and the subsequent examination of how, given those facts, the system impacts the goals and values associated with a particular domain or field of use. (See “Appendix A – Machine Learning Evaluation Framework: Questions.”)

Our starting point is the flow of processing that designers and developers go through in developing Machine Learning systems. Starting with data, and the different forms they take, the ML process then involves algorithmic choices, decisions about functionality, and considerations of the domain in which a system will be deployed. In looking at this process, we want to decouple this last step from the earlier ones in that the domain level considerations are more squarely aligned with evaluation. The core facts of a system remain the same regardless of where it is deployed, but its impact and evaluation are defined by the domain in which it is used. (For more detailed

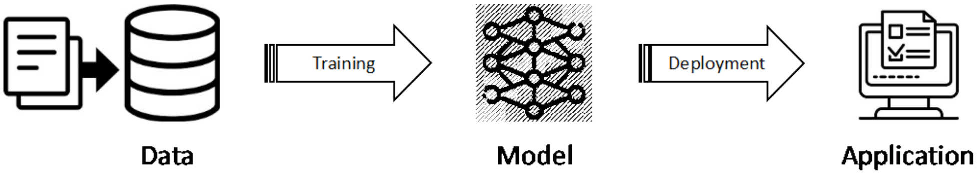


discussion of the machine learning pipeline and development process, see the "Machine Learning Overview and Tutorial" as well as "Appendix B – Machine Learning Algorithm Compendium.")

This division allows us to establish the facts using agreed upon methods. It also isolates where areas of agreement or disagreement exist and reduces the complexity of the process by separating fact from evaluation.

Features and Facts

As we consider applications, we need to tease out the features that are specific to the applications such as data sourcing and quality, the nature of the algorithms used, and the dynamic of user interactions with the resulting system; these considerations parallel the high-level ML development process.



The Basic ML Process Flow

For any given ML application, designers and developers need to move through the same three core phases:

1. Data are gathered, cleaned, normalized, and harmonized against the task at hand and desired learning outcomes.
2. Specific algorithms are selected, specific features of the data are selected, and the model is iteratively trained and tested.
3. User interactions are designed and developed to facilitate functionality and usability.

In each of these phases, designers, developers, and product managers make decisions that impact the performance of the resulting system. The goal of the Framework is to develop a set of questions for each one that uncovers the basic facts of how the system was built and what its expected performance will be.

Evaluation

The process of fact gathering results in a functional description of both the features of an application and the decisions that were made in the process of developing it. With this description in hand, the second stage of the process, domain level evaluation, can proceed.

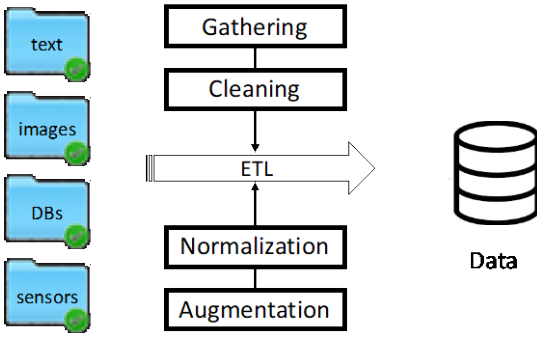
Framework Components

In each of the core component areas, the framework process utilizes a set of questions to interrogate the choices, facts, assumptions, features, constraints and methodologies that were used to develop and provide the structure of the system. (To review the detailed set of Framework questions by component area, see "Appendix A – Machine Learning Evaluation Framework: Questions.")

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 multiple data types (Boehm, Kumar, and Yang 2019). The processes that define data flow also define the set of issues that must be extracted to evaluate the system that results from them. Each of the processes that make up the ecosystem of data collection and integration needs to be examined for quality, completeness and coverage. Each of the steps in the process has its own issues and could impact the performance of any system that is built on top of the models they produce (Polyzotis et al. 2018).

Questions of how the data were gathered, cleaned, integrated with other sources, and augmented all have to be asked to develop a complete set of facts (Carbone, Jensen, and Sato 2016).



Issues in Data Collection

In looking at the original sources, questions of their reliability, coverage, and any user incentives that might skew 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 completeness of examples, and the validity of the sources themselves (Suresh and Guttag 2019).

The data gathering process needs assessed distinctly. Examining how data was gathered, we consider whether the process itself introduces any issues that could compromise completeness or coverage (Khan et al. 2014). Are there features of a data set that are not gathered or examples that are beyond the scope of the ingestion process? Are different data sources ingested in different ways or do they have their features extracted in a manner that pulls them out of alignment?

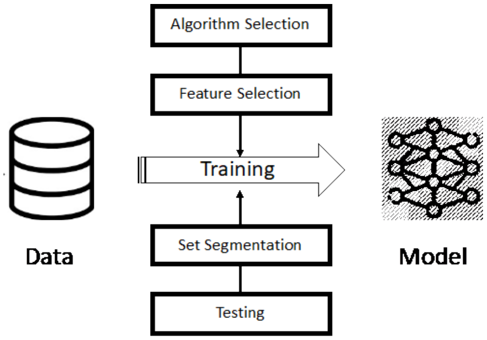
As we consider how an initial data set is processed and enhanced, we need to go through a similar set of queries, here focusing on how specific data elements are extracted (e.g., pulling entries from text), enriched (mapping ambiguous pieces of text onto controlled vocabularies) (Fayyad, Piatetsky-Shapiro, and Smyth 1996), and if new synthetic data sets have been generated to help support the learning that it will drive (Gupta, Vedaldi, and Zisserman 2016).

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 (Kadadi et al. 2014).

These inquiries into the sources and processing of the data are aimed at getting not only the facts as a snapshot of data in their final form but to surface the sources of possible problems so that they can be identified and, if warranted by their impact, remedied.

Algorithmic Choices

Independent of the data are questions related to the algorithms supporting the learning process. While there is some overlap between the analysis of the data and the algorithms that use them, it is important to examine each in its own light. In viewing algorithmic issues, the focus is on two elements: the decisions related to the inputs and training itself, and those related to the features of the models that result. The first of these involves the choices that were made by developers that shape the nature of the model they produce. The second involves the model itself and its performance.



Issues in Model Building

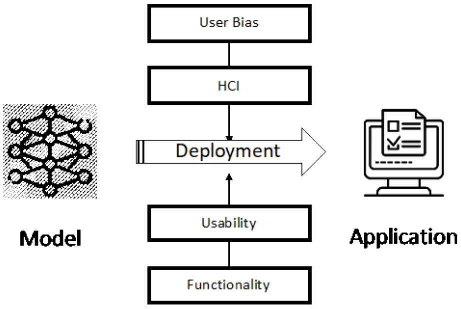
At this stage, developers must make decisions about which algorithms to apply (Ali and Smith 2006), which features may or may not participate (Chandrashekar and Sahin 2014), and how to segment the data set for training and testing (Reitermanova 2010). Developer choices include the features that are used by the system, the cycle of training and testing, the choice of specific algorithm and the training and retraining dynamics. (See “Appendix B – Machine Learning Algorithm Compendium.”) 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. And choices about how a system is updated and retrained can create self-reinforcing predictions (Nguyen et al. 2014).

Looking at the resulting models, we need to consider issues such as their levels of accuracy, or performance as measured by other metrics (Sokolova and Lapalme 2009), and their levels of opacity. While not a value judgement, some systems provide more of a window into their operations and the features that they are utilizing than others (Caravalho, Pereira, and Cardoso 2019). The different levels of transparency impact how and when different models can be utilized.

Different deployments and domains have different requirements (Baier, Jöhren, and Seebacher 2019). Fitting a single ML approach to all of them makes little sense and over-constrains the application of powerful technologies. In order 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.

Interaction

Once a model has been produced, it is incorporated into a larger system. The design of the Human/Computer interaction with the machine learning model has tremendous impact on the ways in which the system can be guided and the ways in which the system guides (Inkpen et al. 2019), which ultimately influence the outcomes and the impact that result. As we examine interactions, we need to ask questions aimed at uncovering the ways in which model results are interpreted and guide user decision-making.



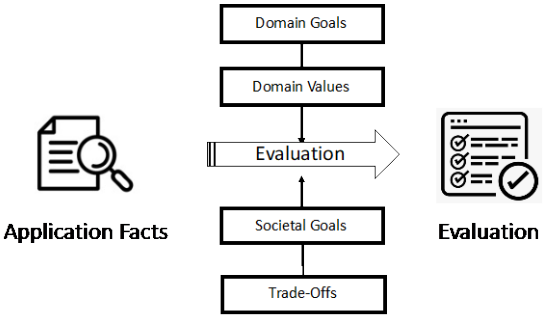
Issues in Application Integration

While these questions require an examination of how humans interact with these systems, they are still in the realm of establishing the facts of the matter. These key questions help to ascertain the likely outcomes of the system in practice, in context. 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 (Vredenburg et al. 2002)? Are they to judge, or even correct, the outputs of the model (Amershi et al. 2014) – or is there a danger of over-trusting those outputs (Kirkpatrick, Hahn, and Haufler 2017)? And what is the role of the user in the system and the nature of the handoff between the machine and the human who is utilizing it?

The goal is not to determine whether the interactions are appropriate but to understand exactly what they are (Inkpen et al. 2019). An interaction involving a handoff from machine to user when the machine is unable to make a decision is a fact of the matter. Whether that handoff can be managed by a user in a given circumstance is a matter of evaluation.

Evaluation

The facts related to the data, the algorithm, and the system interaction are input to the evaluation process. Evaluation is done within the context of the domain in which this application will be utilized. Analyzing an algorithm against the goals and values of a domain equips us with a set of analytical tools to judge the deployment of machine learning as appropriate or inappropriate and to articulate the tradeoffs, benefits, and drawbacks of its human impact.



Evaluation Stage

The domains provide the core goals and values against which to compare the functional facts and the context for reasoning about tradeoffs. This allows the assessment to move beyond generic issues of fairness, accountability, and transparency to consider the specific impact as defined by specific domain-level goals. Additional, more general, societal goals are drawn in using the same structure and mechanisms that the various domains employ themselves.

In evaluating the propriety of a machine learning system in the context of its domain goals and values, there are six key elements to consider: the Primary Goals of the application, impact on Secondary Goals, impact on Implicit or Background Goals, possible Negative Impact on Individuals, possible Negative Impact on Groups, and possible Negative Impact over Time.

Goals

Institutions, professions, and (loosely) social contexts — what we are clustering together as "domains" — all have goals (Walzer 2008). 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 these goals as impact requirements that must be met for a system to be acceptable (Van de Poel 2013; Richardson 1997). We defined a goal as "an outcome we hope to accomplish in an institution, profession or social context."

In this situation, "goal," is used aspirationally as opposed to descriptively. We are not trying to describe people's actual motivations, in that they might be motivated by fame, reputation, money, vengeance or any number of personal aims outside of domain. We are interested in augmenting and catalyzing the positive contributions that these domains make to society. And each of these domains, such as journalism, medicine, or business, has some positive, characteristic benefit that it supposes itself to make to society.

Even domain practitioners might have difficulty articulating the goals and values of their domain let alone considering the impact of an automated system with regard them. With that in mind, we developed several prompts to help practitioners identify the goals of their domain. Ideally, these prompts would converge on one or a set of general answers. In some cases, there may be empirical evidence that these goals are endorsed by the profession, e.g., statements from professional organizations, or professional codes of ethics.

To identify the goals of a domain, it is helpful to ask:

- Why do people choose to go into this field over others?
- What are people within this domain hoping to contribute to society? What do they take their *raison d'être* to be?
- How do the people working in this domain praise themselves, e.g. in their advertisements, award ceremonies, or public statements?
- What is the benefit that is peculiar to this institution, that is not provided by other institutions in society?
- What benefits do consumers, users, or broader society expect these institutions to furnish? What is the point of these domains in the eyes of outsiders?

This list is valuable for identifying the core goals and values of domain and establishing a set of target requirements. This provides us with the requirements that can now be used to test the utility of an application by considering how well the facts and performance of the system satisfy those requirements.

With the facts of a system and domain-level requirements, we can identify potential mis-alignment between the goals of a domain and the functional performance of a machine learning system. For example, consider the core purpose of the application and whether the model output could be optimizing to a proxy value of what is desired. Might that proxy fail to reflect the genuine goal of the domain?

And when a domain has multiple goals, 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. Are there goals associated with the task that are different from the goals that the system is focused on? Are there any goals outside of the focus area of application that are impacted by its utilization? Similarly, some effects of a model might not manifest in the immediate term, but only over time. Could using this system as a long-term policy distort the functioning of the institution or domain it’s deployed into?

Values

There is some nuance to the distinction between goals and values. We defined values 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).

In our usage, a value is a feature of our actions that is important to us. 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. To identify the values of a domain, it is helpful to consider what kinds of actions are criticized or punished, either legally or by the censure of one's colleagues. What kind of behavior is seen as unbecoming?

To put these definitions together: goals define what an application is designed to accomplish. Values define the issues that need to be considered as the application is doing so.

These domain-level requirements include goals and values that go well beyond the application. A system that is able to use images to provide on-point identification of renal tumors has a set of very specific application-level requirements related to accuracy and precision. But the goals associated with medicine in general, such as “Do no harm,” and societal goals driven by fairness and access are also part of the consideration of goals and values of the domain.

There is one issue worth discussing here, which is a cluster of concerns about the theoretical viability of attributing goals and values to institutions. 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 (e.g. within the film industry, compare the goals of producers, directors, and writers). Third, individuals within a domain might have different goals than the goals we attribute to the domain itself. The people inside an institution who are answering phones or processing invoices might not have any particularly lofty goals at all. All of these objections express skepticism that we can treat domains or professions 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 professions or institutions purport to provide. This is best seen as a metaphor, so that we do not have to ascribe desires or intentions to institutions. We are also confident we could identify the goals that most people in the profession would agree upon.

Applying the Framework

The Framework is designed to help guide the deconstruction of applications in order evaluate their potential impact on both human and societal health and safety. While the Framework directs focus to the components, not all problematic applications will have issues with all components. The questions remain the same, but not all answers will uncover issues. Simply put, an application with data issues may be problematic but that does not mean that either the algorithm used in the learning or the design of the interactions have issues as well.

For example, during the pandemic, many hospitals piloted computer vision algorithms that analyzed lung scans from patients with the novel coronavirus and were used to train models used in categorization (Wehbe et al. 2021; Castiglioni et al. 2020; Wang, Lin, and Wong 2020). The goal of these systems was to diagnose possible COVID patients and to reduce the turnaround time of diagnosis.

In follow-up work, researchers found that the chest X-ray datasets used to train some of these models exhibited imbalance, biasing them against some gender, socioeconomic, and racial groups (Seyyed-Kalantari et al. 2020). As a result, the models under-diagnosed individuals in groups that were underrepresented. The research uncovered problems related to the data, though, from the perspective of the algorithmic choices and ways in which physicians interacted with the systems, there may not have been any issues. Interestingly, we have heard anecdotes that once the initial research results were made known, groups working on image-based COVID diagnosis reexamined their data from the perspective of population distribution, which led many to either withdraw systems or expand their data sets to attempt to rectify this imbalance.

In the following sections, we examine some use cases that demonstrate issues that can arise from each of the Framework components.

Data Issues

The initial framework questions used to examine any ML application are focused on data.

In the COVID diagnostic example, the problem was that the distribution of examples in the data set did not reflect the actual distribution of instances in the world. While distribution is one of the more common issues in data, it is, by no means, the only one.

In looking at data gathering, the Framework question of whether there are impediments to accurate data collection is rarely asked. But when it is, we can uncover problems in a wide variety of scenarios and circumstances. In crime data, researchers uncovered problems in the collection process that flowed from the nature of how charges against an individual are decided upon (National Research Council

2003). Charges are based upon what can be proven, not what someone has done. These changes are reflected in the data, data that was then used to train a model and then build predictive policing systems. The actual impact of this data problem remains an open question.

When data are gathered from different sources, coordination problems can arise (Halevy, Rajaraman, and Ordille 2006; Howard et al. 2021). For example, in the development other COVID diagnostic systems, examples of images were collected from multiple sources. While the system did well against initial test examples, it failed as researchers discovered that images taken from some sources shared orientation and labeling similarities that had become tied to diagnoses. The system, in effect, learned to recognize that examples were from particular sources that happened to consist of a preponderance of COVID cases (DeGrave, Janizek, and Lee 2021).

While corrected in this instance, it has not led to the adoption of best practice regarding data integration and correction for spurious correlations. The simple question of whether there are possible confabulating features in one data set or another still tends to be unasked. And, without problems that might be surfaced in the application, they may never be identified. Unfortunately, errors like this often result in excellent performance numbers in testing that push developers in the direction of adoption.

For data, the questions range from the way the data are collected, choices of sources, features selection, the veracity of the data, the methods used to integrate multiple data sources, and methods used to augment it. And even when the individual errors are small, they can amplify each other. A series of compromises that drop performance by 5% each, quickly drops the utility of the data in a matter of four or five steps. But with a focus on unit testing, these incremental degradations often go unnoticed even after deployment.

Algorithmic Issues

When most people consider the problems associated with systems based on Machine Learning, they tend to use the phrase “algorithmic bias” (Danks and London 2017). The reality is that while there are many problems that can be introduced into a system because of algorithmic decisions, it is only one of the places where this occurs, and “bias” is only one of the real problems that these systems exhibit.

From the Framework perspective, the starting point is the choice of algorithms themselves. In general, there is a tradeoff between the precision and accuracy of an algorithm and our ability to understand the model that results from it (Adadi and Berrada 2018). A system built on top of Naïve Bayes may not perform as well as one utilizing a model constructed using Recurrent Neural Nets but is far more transparent than its more precise counterpart. (See “Appendix B – Machine Learning Algorithm Compendium.”)

If we consider downstream utilization and requirements (e.g., use in highly regulated industries, need for auditing, application level considerations about bias or fairness), the transparency of the model can be tremendously important. Systems built on top of opaque system can only be tested post hoc using manual techniques to compare the system’s results against some other external standard. For example, consider the COMPAS system, about which ProPublica published a bombshell investigative story in 2016 which galvanized the public conversation around the fairness, accuracy, and transparency of algorithms (Angwin et al. 2016). Because the core algorithm used in COMPAS was considered proprietary, the only methods available to evaluators were extensive manual testing of input and output behaviors (Larson et al. 2016). In this instance, the testing was necessary in that the system itself had massive impact on human life. In other areas such as product recommendations, the need for this kind of deep dive may be less pressing. But in both deployment instances the facts regarding algorithmic decisions and transparency remain the same.

Similarly, as systems are trained, designers and developers make decisions about which features of a data set should be included (Chandrashekar and Sahin 2014), which data sets are going to be part of the training (Roh, Heo and Whang 2019), and how the data are segmented into training and testing corpora (Reitermanova 2010). Each of these decision areas impacts the use of systems employing the resulting model, potentially to ill effect.

For example, feature selection, used to reduce training time and remove either redundant or irrelevant features, can result in removing independent variables that, in turn, skew results. In medicine, the removal of meta data identifying sources of examples can weaken results in that different hospitals may have different diagnostic thresholds or techniques which makes transferring learning from one data set to another difficult.

In a recent Q&A session[1], Andrew Ng, noted ML leader, commented on this issue:

> *when we collect data from Stanford Hospital, then we train and test on data from the same hospital, indeed, we can publish papers showing [the algorithms] are comparable to human radiologists in spotting certain conditions. It turns out [that when] you take that same model, that same AI system, to an older hospital down the street, with an older machine, and the technician uses a slightly different imaging protocol, that data drifts to cause the performance of AI system to degrade significantly.*

Likewise, efforts to remove meta-data such as race or ethnicity from examples, in service of removing bias, can introduce error into the model. A study published in 2020 showed that including race data in medical records and using it to correct for existing inequalities improved overall performance and significantly reduced bias (Allen et al. 2020).

Testing segmentation – dividing a data set into a corpus for training and a corpus for testing – is often done iteratively with the goal of developing the most accurate results. Unfortunately, this can result in testing corpora that are populated by the clearest of cases in that they provide the clearest of results. In applications that have significant edge cases – medicine, the law, hiring decisions – this results in systems that miss all but the obvious decisions. By surfacing the techniques used to determine the training/test sets, we surface the areas in which there may be problems.

Likewise, as developers train systems they make decisions about which data sets to include or exclude. The effect of this is that some data that provided the appropriate coverage in a space is removed from the training, or data that augmented the features in a core set are not incorporated. In many cases, this is intended to remove outliers or noise in the data but has the outcome of narrowing the scope of

[1] May 3rd, 2021 reported by IEEE Spectrum. https://spectrum.ieee.org/andrew-ng-xrays-the-ai-hype Accessed 9/2/2021

effectiveness of the model. Specific drug treatment policies at one hospital, if removed from the data to generalize the corpus, can have the effect of making the model inapplicable at other hospitals where the policies are different (Futoma et al. 2020).

As with many decisions, this may be less about the immediate impact and more understanding that the features excluded may have been able to provide more nuance or precision; their exclusion and the reasons why are part of the landscape of facts that will participate in both evaluating systems and subsequently correcting them.

Interaction Issues

The dynamic of how systems interact with users is often ignored in the consideration of their impact. But the reality is that the details of interaction between intelligent systems and human users can both introduce problems in systems that are otherwise benign and compensate for biases that might be inherent in the system.

For real-time systems in which control is shared between a human user and a system, we must understand the nature of the hand off and how context is maintained. For example, when we look at self-driving cars that maintain control under most circumstances but hand that control over to the driver under difficult conditions, the question of how to communicate (or maintain) a driver's understanding of the situation is paramount. Having drivers reestablish context is simply not feasible in real-time (and often emergency) situations and, as we have seen, can lead to deadly outcomes (Griggs and Wakabayashi 2018). In evaluating such systems, we must first establish the mechanisms for trade-off (time frame, context sharing, warning mechanisms) and then consider the ways in which user attention can be maintained during the process.

A set of somewhat more subtle issues arise from the problem of loss of skills because of reliance on automated systems. As was discovered with auto-pilot systems, human pilots' skills tended to grow stale because of lack of engagement. This has resulted in situations in which the machine ceded control of the plane to human users whose skills were not sharp enough to respond appropriately (Oliver, Calvard, and Potočnik 2017). For the aviation industry, this problem is mitigated through training maintenance requirements, but it is not clear (and an open research question) how to maintain those skills for every driver sitting in the seat of a self-driving car as they roll out.

For decision support systems, we need to focus on a different set of questions primarily focused on the role of support. In medicine, the view is that diagnostic systems provide "suggestions" that a physician uses to incorporate into a broader diagnostic picture. Unfortunately, physicians are human and their reliance on the machine's suggestions can drift toward seeing the machine's diagnosis as ground truth. As with the hand-off of control in self-driving cars, an approach to the problem from the perspective of policy alone is not enough. We must consider how to either enforce that policy or provide mechanisms at the interaction level to explicitly integrate the machine's output with other features.

In a similar view, the framing of results shifts the focus of how users incorporate them. The same underlying technology system can be used to suggest possible solutions that a user can incorporate into their decision-making process or be used to critique solutions proposed by the user. In the former, the machine's solution becomes the starting point that may or may not be modified. It becomes the default answer. In the latter, the user's solution becomes the default that may or may not be changed in light of the machine's comments. No matter how accurate the automated solution is, the role that it plays in the decision dynamic can change how it impacts the decision-making process and outcomes.

Decisions about interactions can have dramatic consequences but are often overlooked in the deployment of ML systems. The framing of questions as either opt-in or opt-out – for example, in organ donation questions during driver's license renewal – can determine outcomes more than almost any other feature. In the organ donor case, the question framed as an opt-in ("Check here if you want to be an organ donor.") nets on the order of 42% participation while the same question framed as an opt-out ("Check here if you do not want to be an organ donor.") nets 82% participation (Johnson and Goldstein 2003). The dynamic of these interactions and how decisions are managed are crucial in determining outcomes and require us to ask questions as to how they are managed.

User attitude toward machine recommendations is impacted by the level of understanding that is provided to users (e.g., explanations, alternatives, trade-offs) as well as control. In the former, answers that are simply that, answers with no rationale, can either be ignored or adopted on faith. With explanations, users are presented with more information about the basis of the recommendation that can then be used to evaluate it. In the latter, the ability to change inputs or context create a different relationship between the machine and its users. The outputs become responses to hypotheticals rather than unquestioned answers. The experience provokes more rather than less thought.

There is an irony here. While industry has developed a set of what tend to be called "dark patterns" – interactions that are designed to addict or manipulate users (e.g., teaser recommendations, gamification of decision-making, framing) – consideration of the impact of these sorts of interaction design decisions is kept at arm's length by the human-computer interaction community. As a result, they are used – primarily by organizations that are attempting to use them for commercial ends. Regardless of intent, however, understanding the design and consequences of these interactions is crucial to considerations of human health and safety.

Goals and Values Alignment Issues

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, again consider the case of COMPAS. ProPublica reported that a company in Florida, NorthPointe, developed an algorithm called "Correctional Offender Management Profiling for Alternative Sanctions" (COMPAS) for assessing the risk that someone convicted of a crime would reoffend, based on 100+ factors (Angwin et al. 2016). 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 in a specific way: the algorithm tended 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, but is usually thought to be one of rehabilitating and reintegrating former prisoners (Lynch 2000; Simon 1993) by releasing 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 its goal — even if the model's predictions were perfect.

Consider next a comparison between two uses of recommendation systems. These systems recommend things to the user that seem relevant to them given their past behavior or profiles of users like 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, precisely because it aligns with what the domain of advertising contributes to society.

However, that use of machine learning becomes problematic when the same algorithm is used to curate the information users see on their Facebook newsfeed. There is a significant difference between the goals of advertising and the goals of journalism and it is unlikely that a single algorithmic approach to information would serve both. The purpose of journalism is at least in part to tell us things that *we need to know*, even if we don't *want* to know them or don't *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're being deployed — even if the same basic system is used in each case.

Finally, consider a more recent news story about using machine learning to predict students' scores on a college entry exam (Satariano 2020). College-bound students in England take a standardized test called the Advanced Level (or "the A level"), which is influential in determining their competitiveness for universities. In 2020, those tests were canceled because the spread of COVID-19 made testing in person unsafe. Instead, the British government used a machine learning model to predict what students *would* have scored if they *had* sat for the exams. As a result, around 40% of students saw their grades — along with the overall competitiveness of their college application — fall, and many students had their college admissions offers rescinded.

This seems problematic, again, even if we imagine that this system were perfect at predicting what students *would have* scored on the A level exam. This use of machine learning seems to pervert the functioning of the educational institution. The goals of higher education are debatable, of course, but some plausible goals include: to impart skills to students that will help them live good lives, such as discipline, time management, and preparation; or to nurture students' development into responsible adults and citizens of a dynamic world. Meanwhile, the values of education include things like: equality of opportunity, fairness, second chances, and merit. We want the opportunities that students have to align with their abilities and their mastery of the material. This use of machine learning is inappropriate, then, because it doesn't align with these plausible goals and values of education.

It is worth considering an objection that several workshop participants raised to this last example. The problem here, a critic might point out, is not primarily with machine learning, but with standardized testing itself. Here, the concerns over automating some parts of the educational institution through machine learning point to a deeper criticism of the pre-existing practices of the institution. The use of machine learning in this case is problematic, but the *prior* use of standardized testing also could violate the goals and values of education. We do not take this to undermine the general strategy of examining domain goals and values. If anything, the exercise can help us illuminate preexisting practices that might violate those goals and values — which the introduction of machine learning could exacerbate. In this case, that was revealed to be the use of standardized testing metrics in the education system in general, which of course predates the use of machine learning.

Research Vision, Roadmap and Next Steps

Our goals in developing this Framework are two-fold.

First, we intend to develop a set of best-practices that could be used in the evaluation and ultimately development of Machine Learning applications. The goal is a repeatable and operational process for the identification of the potential negative impacts of machine learning applications and the causes of those impacts.

Our initial workshop was used to refine the Framework design by testing it against specific real-world cases. Participants with domain expertise applied and gave feedback on the process to help to refine the Framework. (See the “Machine Learning Impact Initiative Activity Summary” document for more details on the workshops.) The result was mapped onto the structure outlined in this document, a Framework designed to be used both before and after application development and deployment. The Framework was defined, refined, and tested with this goal in mind.

Second, we utilized the Framework to identify areas where presently it is difficult or impossible to answer the specific detailed questions in the framework, to anticipate or prevent issues represented by the questions, or to test for them. These areas are key opportunities where research is needed to fill in the gaps in the gathering of facts related to the core development areas (data, algorithms, interaction) and the issues related to evaluation. The goal was, and continues to be, the identification of gaps discovered through the application of the Framework to specific systems to determine when and where there might be problems applying it. Our goal is to identify problems that exist because of lack of knowledge about the details of applications and the processes used to develop them; problems that point to research that needs to be done.

The second MLII workshop was used to identify these research areas utilizing experts in the specific Framework component areas (data, algorithms, interaction, evaluation). Participants were tasked to explore examples and identify specific information issues that require research to solve. The identified issues were incorporated into the Research Roadmap. The Roadmap defines a set of initial specific research problems that are necessary in order to further operationalize the design, development, and evaluation of AI systems from the perspective of human health and safety. (See the "Roadmap for Research on the Human Impact of Machine Learning Applications" document for the consolidated results of these efforts.)

Both the Framework and the Roadmap are intended to evolve. In their current states, they provide us with a starting point for working through approaches and research that will allow us to develop a set of practices aimed at improving the effectiveness of ML applications and assuring that they are designed to function without threatening human health and safety.

At its base, the Framework provides a foundational structure for the design and evaluation of Machine Learning applications. It also provides us with a touchpoint for defining a Research Roadmap for work needed to operationalize it. Utilizing the key questions and concerns of each component, we established the core platform and used it to establish a central set of important research problems.

As we move forward, we have three primary thrusts to future work on the Framework itself: application, refinement, and road mapping.

> **Application**: While the Framework has theoretical validity, it needs to be tested using real world applications. The testing that was done at the workshop level gave us important initial information; the next step is to disseminate the model and track how it is used and how effective it is in guiding the evaluation process.
>
> **Refinement**: As with any approach to best-practice, the Framework remains a work in progress. Our goal in the near term is to continue to refine the model in response to any issues that arise in its utilization. Identifying these issues can result in either refinement or transformation of the Framework itself or establishment of a research plan to develop approaches that are needed to fill in information gaps.
>
> **Road Mapping**: While some issues that arise in the application of the Framework may lead to refinement of the model, often the result is the realization that certain core facts associated with a system – either while being evaluated or developed – might not be readily available. Methods for uncovering these facts are now possible research problems that can be investigated by the research community.

As we move this approach forward, the goal is to provide the community with an evolving set of tools that both evaluators and developers can use to operationalize assessment of concerns about the impact of Machine Learning systems on human health and safety.

Citations

Adadi, Amina, and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)." *IEEE access* 6 (2018): 52138-52160.

Ali, Shawkat, and Kate A. Smith. "On learning algorithm selection for classification." *Applied Soft Computing*, *6*(2), (2006): 119-138.

Alikhademi, Kiana, Emma Drobina, Diandra Prioleau, Brianna Richardson, Duncan Purves and Juan E. Gilbert. A review of predictive policing from the perspective of fairness. Artif Intell Law (2021). https://doi.org/10.1007/s10506-021-09286-4

Allen, Angier, Samson Mataraso, Anna Siefkas, Hoyt Burdick, Gregory Braden, R Phillip Dellinger, Andrea McCoy, et al. "A Racially Unbiased, Machine Learning Approach to Prediction of Mortality: Algorithm Development Study." JMIR public health and surveillance vol. 6,4 e22400. 22 Oct. 2020, doi:10.2196/22400

Allen, Colin, Iva Smit, and Wendell Wallach. "Artificial morality: Top-down, bottom-up, and hybrid approaches." Ethics and information technology 7.3 (2005): 149-155.

Amershi, Saleema, Maya Cakmak, William Bradley Knox, and Todd Kulesza. "Power to the people: The role of humans in interactive machine learning." *AI Magazine*, *35*(4), (2014): 105-120. https://doi.org/10.1609/aimag.v35i4.2513

Anderson, Elizabeth. *Value in ethics and economics*. Cambridge, MA: Harvard University Press, 1993.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." *ProPublica*, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. May 23, 2016. Accessed 28 May 2021.

Baier, Lucas, Fabian Jöhren, and S. Seebacher. "Challenges in the deployment and operation of machine learning in practice." *ECIS (2019).*

Boehm, Matthias, Arun Kumar, and Jun Yang. *Data Management in Machine Learning Systems. Synthesis Lectures on Data Management*, *11*(1), (2019).

Brey, Philip. "Ethical aspects of facial recognition systems in public places." Journal of information, communication and ethics in society (2004).

Carbone, Anna, Meiko Jensen, and Aki-Hiro Sato. "Challenges in data science: A complex systems perspective." *Chaos, Solitons & Fractals*, *90*, (2016): 1–7. https://doi.org/10.1016/j.chaos.2016.04.020

Carvalho, Diogo V., E. M. Pereira and Jaime S. Cardoso. “Machine Learning Interpretability: A Survey on Methods and Metrics.” Electronics 8 (2019): 832.

Castiglioni, Isabella, Davide Ippolito, Matteo Interlenghi, Caterina Beatrice Monti, Christian Salvatore, Simone Schiaffino, Annalisa Polidori, Davide Gandola, Cristina Messa, and Francesco Sardanelli. "Artificial intelligence applied on chest X-ray can aid in the diagnosis of COVID-19 infection: a first experience from Lombardy, Italy." *MedRxiv* (2020).

Chandrashekar, Girish and Ferat Sahin. "A survey on feature selection methods." *Computers & Electrical Engineering*, *40*(1), (2014):16-28.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, and Sharad Goel. “A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It’s Actually Not That Clear.” *Washington Post*. April 18, 2019. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/. Accessed 5/28/2021.

Dancy, J. "Should we pass the buck?" In T. Rønnow-Rasmussen & M. J. Zimmerman (Eds.), *Recent Work on Intrinsic Value.* Dordrecht: Springer. (2005): 33–44.

Danks, David, and Alex John London. "Algorithmic Bias in Autonomous Systems." In *IJCAI*, vol. 17, (2017): 4691-4697.

Davis, Michael, Andrew Kumiega, and Ben Van Vliet. "Ethics, finance, and automation: A preliminary survey of problems in high frequency trading." *Science and engineering ethics* 19.3 (2013): 851-874.

DeGrave, Alex J., Joseph D. Janizek, and Su-In Lee. "AI for radiographic COVID-19 detection selects shortcuts over signal." *Nat Mach Intell* 3, 610-619 (2021). https://doi.org/10.1038/s42256-021-00338-7

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." *AI magazine*, *17*(3), (1996):37-37. https://doi.org/10.1609/aimag.v17i3.1230.

Futoma, Joseph, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. "The myth of generalisability in clinical research and machine learning in health care." The Lancet Digital Health. Vol 2 Issue 9 (2020): e489-e492. https://doi.org/10.1016/S2589-7500(20)30186-2.

Goodman, Bryce, and Seth Flaxman. "European Union regulations on algorithmic decision-making and a “right to explanation”." *AI magazine 38(3),* (2017): 50-57.

Griggs, Troy and Daisuke Wakabayashi. "How a Self-Driving Uber Killed a Pedestrian in Arizona." The New York Times. https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html Accessed 9/2/2021.

Gupta, Ankush, A. Vedaldi, and Andrew Zisserman. "Synthetic Data for Text Localisation in Natural Images." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016): 2315-2324.

Halevy, Alon, Anand Rajaraman, and Joann Ordille. "Data integration: The teenage years." In *Proceedings of the 32nd international conference on Very large data bases*, (2006): 9-16.

Howard, Frederick M., James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo et al. "The impact of site-specific digital histology signatures on deep learning model accuracy and bias." *Nature communications*12, no. 1 (2021): 1-13.

Inkpen, Kori, Stevie Chancellor, Munmun De Choudhury, Michael Veale, and Eric P. Baumer. "Where is the Human? Bridging the Gap Between AI and HCI." *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI EA '19) (2019):1-9.

Johnson, Eric and Daniel Goldstein. "Do Defaults Save Lives?" Science, Vol. 302, Issue 5649: 1338-1339. (2003)

Kadadi, A., R. Agrawal, Christopher Nyamful, and Rahman Atiq. "Challenges of data integration and interoperability in big data." *2014 IEEE International Conference on Big Data (Big Data)* (2014): 38-40.

Kaminski, Margot E. "The right to explanation, explained." *Berkeley Tech. LJ* 34 (2019): 189.

Khan, Nawsher, Ibrar Yaqoob, I. A. T. Hashem, Zakira Inayat, Waleed Kamaleldin Mahmoud Ali, Muhammad Alam, M. Shiraz and A. Gani. "Big Data: Survey, Technologies, Opportunities, and Challenges." *The Scientific World Journal 2014*. (2014)

Kirkpatrick, Jesse, Erin N. Hahn, and Amy J. Haufler. "Trust and Human-Robot Interactions." *Robot Ethics 2.0: from Autonomous Cars to Artificial Intelligence* (eds. Patrick Lin, Ryan Jenkins and Keith Abney) (2017).

Kroes, Peter, Maarten Franssen, Ibo van de Poel, and Maarten Ottens. "Treating socio-technical systems as engineering systems: some conceptual problems." Systems Research and Behavioral Science: The Official Journal of the International Federation for Systems Research 23.6 (2006): 803-814.

Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. “How We Analyzed the Compas Recidivism Algorithm.” ProPublica. ProPublica, May 23, 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed on 9/2/2021.

Lynch, Mona. "Rehabilitation as rhetoric: The ideal of reformation in contemporary parole discourse and practices." *Punishment & Society* 2.1 (2000): 40-65.

Max, Raphael, Alexander Kriebitz, and Christian Von Websky. "Ethical Considerations About the Implications of Artificial Intelligence in Finance." Handbook on Ethics in Finance (2020): 1-16.

Mesthene, Emmanuel G. "The role of technology in society." *Technology and values* (ed. Kristin Schrder-Frechette) (1997): 71-85.

National Research Council. Measurement Problems in Criminal Justice Research: Workshop Summary. Washington DC: The National Academies Press. 2003. https://doi.org/10.17226/10581

Nguyen, Tien T., Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. "Exploring the filter bubble: the effect of using recommender systems on content diversity." *Proceedings of the 23rd international conference on World wide web (WWW '14)* Association for Computing Machinery, New York, NY (2014): 677-686.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting racial bias in an algorithm used to manage the health of populations." Science, Vol. 366, Issue 6464: 447-453. (2019)

Oliver, Nick, Thomas Calvard, and Kristina Potočnik. "The Tragic Crash of Flight AF447 Shows the Unlikely but Catastrophic Consequences of Automation." Harvard Business Review. https://hbr.org/2017/09/the-tragic-crash-of-flight-af447-shows-the-unlikely-but-catastrophic-consequences-of-automation. Accessed 9/2/2021.

Polyzotis, Neoklis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. "Data Lifecycle Challenges in Production Machine Learning: A Survey." *ACM SIGMOD Record*, *47*(2), (June 2018):17–28. https://doi.org/10.1145/3299887.3299891

Raji, Inioluwa Deborah, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. "Saving face: Investigating the ethical concerns of facial recognition auditing." In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, (2020): 145-151.

Reitermanova, Z. "Data Splitting." *WDS'*10. (June 2010): 31-36.

Richardson, Henry S. *Practical reasoning about final ends*. Cambridge University Press, 1997.

Rodgers, Shelly. "Themed issue introduction: Promises and perils of artificial intelligence and advertising." *Journal of Advertising*. (2021): 1-10.

Roh, Yuji, Geon Heo, and Steven Euijong Whang. "A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective." *IEEE Transactions on Knowledge & Data Engineering* 01 (2019): 1-1.

Satariano, Adam. "British Grading Debacle Shows Pitfalls of Automating Government." *The New York Times*, 20 Aug. 2020. NYTimes.com, https://www.nytimes.com/2020/08/20/world/europe/uk-england-grading-algorithm.html.

Selbst, Andrew, and Julia Powles. "“Meaningful Information” and the Right to Explanation." Conference on Fairness, Accountability and Transparency. PMLR, (2018).

Selinger, Evan, and Brenda Leong. "The ethics of facial recognition technology." Forthcoming in *The Oxford Handbook of Digital Ethics* (ed. Carissa Véliz) (2021).

Seyyed-Kalantari, Laleh, Guanxiong Liu, Matthew McDermott, Irene Y. Chen, and Marzyeh Ghassemi. "CheXclusion: Fairness gaps in deep chest X-ray classifiers." In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, (2020): 232-243.

Simon, Jonathan. *Poor discipline*. University of Chicago Press, 1993.

Sokolova, Marina, and Guy Lapalme. "A systematic analysis of performance measures for classification tasks." *Information processing & management*, *45*(4), (2009): 427-437. https://doi.org/10.1016/j.ipm.2009.03.002.

Suresh, Harini and J. Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle." *arXiv preprint arXiv:1901.10002*. (2019)

Van de Poel, Ibo. "Translating values into design requirements." *Philosophy and engineering: Reflections on practice, principles and process*. Springer, Dordrecht. (2013): 253-266.

Van de Poel, Ibo. "Embedding Values in Artificial Intelligence (AI) Systems." *Minds and Machines* 30.3 (2020): 385-409.

Vredenburg, Karel, Ji-Ye Mao, Paul W. Smith, and Tom Carey. A survey of user-centered design practice. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '02) (2002): 471-478. https://doi.org/10.1145/503376.503460

Walzer, Michael. *Spheres of justice: A defense of pluralism and equality*. Basic books, 2008.

Wang, Linda, Zhong Qiu Lin, and Alexander Wong. "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images." *Scientific Reports* 10, no. 1 (2020): 1-12.

Wehbe, Ramsey M., Jiayue Sheng, Shinjan Dutta, Siyuan Chai, Amil Dravid, Semih Barutcu, Yunan Wu et al. "DeepCOVID-XR: an artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large US clinical data set." *Radiology* 299, no. 1 (2021): E167-E176.

Wieringa, Maranke. "What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. (2020).

Wilcox, Clair. "Parole: Principles and Practice." *Am. Inst. Crim. L. & Criminology* 20 (1929): 345.

Zhang, Daniel, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, et al. "The AI Index 2021 Annual Report," AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA. (2021)