RC25626 (WAT1610-018) October 5, 2016
Computer Science

IBM Research Report

On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products

Kush R. Varshney, Homa Alemzadeh
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598 USA

On the Safety of Machine Learning:
Cyber-Physical Systems, Decision Sciences, and Data Products

Kush R. Varshney and Homa Alemzadeh
Science and Solutions
IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598
Email: {krvarshn, alemzadeh}@us.ibm.com

Abstract—**Machine learning algorithms are increasingly influencing our decisions and interacting with us in all parts of our daily lives. Therefore, just like for power plants, highways, and a myriad of other engineered socio-technical systems, we must consider the safety of systems involving machine learning. Heretofore, the definition of safety has not been formalized in the machine learning context; in this paper, we do so by defining machine learning safety in terms of risk, epistemic uncertainty, and the harm incurred by unwanted outcomes. We then use this definition to examine safety in all sorts of applications in cyber-physical systems, decision sciences and data products, finding that the foundational principle of modern statistical machine learning, empirical risk minimization, is not always a sufficient objective. In particular, we note an emerging dichotomy of applications: ones in which safety is important and risk minimization is not the complete story (we name these Type A applications), and ones in which safety is not so critical and risk minimization is sufficient (we name these Type B applications). Finally, we discuss how four different strategies for achieving safety in engineering (inherently safe design, safety reserves, safe fail, and procedural safeguards) can be mapped to the machine learning context through interpretability and causality of predictive models, objectives beyond expected prediction accuracy, human involvement for labeling difficult or rare examples, and user experience design of software and open data.**

I. Introduction

In recent years, machine learning algorithms have started influencing every part of our lives, including health and wellness, law and order, commerce, entertainment, finance, human capital management, communication, transportation, and philanthropy. As the algorithms, the data on which they are trained, and the models they produce are getting more powerful and more ingrained in society, questions about *safety* must be examined. It may be argued that machine learning systems are simply tools, that they will soon have a general intelligence that surpasses human abilities, or something in-between, but from all perspectives, they are technological components of larger socio-technical systems that may have to be engineered with safety in mind [1].

Safety is a commonly used term across engineering disciplines connoting the absence of failures or conditions that render a system dangerous [2], cf. safe food and water, safe vehicles and highways, safe medical treatments, safe toys, safe neighborhoods, and safe industrial plants. Each of the domains has specific design principles and regulations that are applicable only to it; only a few works in the literature attempt a precise definition applicable to a broad set of domains and systems [3].

In particular, a general definition of safety is the minimization of *risk* and *epistemic uncertainty* (understood in the usual decision-theoretic senses of the words) associated with unwanted outcomes that are severe enough to be seen as *harmful* [3]. The epistemic uncertainty part of the definition is key, because harmful outcomes often occur in regimes and operating conditions that are rare, unexpected, or underdetermined. The cost magnitude of unwanted outcomes is also key, because safety is not concerned with reducing undesired outcomes of an inconsequential nature.

With such a definition of safety, it is possible to consider domains that do not have existing safety principles and regulations such as machine learning [4]. The first contribution of this work is to critically examine the foundational statistical machine learning principles of empirical risk minimization and structural risk minimization [5] from the perspective of safety. We discuss how they, as their names imply, do not deal with epistemic uncertainty. Furthermore, the principles rely on average losses and laws of large numbers-type arguments, which may not necessarily be fully applicable when considering safety. Moreover, the loss functions involved in the formulations are abstract distortions between true and predicted values rather than application-specific quantities measuring loss of life, loss of quality of life, etc. that can be judged harmful or not [6]. To the best of our knowledge, there is no existing work on analyzing machine learning using precise decision-theoretic definitions of safety except our own preliminary work [4].

A second contribution of this paper emerges from examining safety in formulating machine learning problems. Today, machine learning technologies are being used in a variety of settings, including cyber-physical systems, decision sciences, and data products. By cyber-physical systems, we mean engineered systems that integrate computational algorithms and physical components [7]; by decision sciences, we mean the use of algorithms to aid people in making important decisions and informing strategy [8]; and by data products, we mean the use of algorithms to automate informational products such as search and recommendation [8]. These settings vary widely in terms of their interaction with people, scale of data, time scale of operation and consequence, and cost magnitude of consequence. A further contribution is a discussion on how to even understand and quantify the desirability and undesirability of outcomes along with their costs. To complement simply eliciting such knowledge directly from people, we suggest a data-driven approach for characterizing harms that is particularly relevant for cyber-physical systems with large state spaces of outcomes.

Based on these factors, we find that applications of machine learning systems cluster into two types: (A) applications of high consequence that can have a profound effect on people's lives in a short time, and (B) applications of low consequence, usually at a very large scale. Type A applications are the ones in which safety is paramount. We have previously noted the dichotomy of Type A and Type B applications of machine learning and data science in [9], but did not pose them as consequences of safety definitions. The related literature is cited in [9], but again, does not stem from safety.

The final contribution of the paper is a discussion of strategies to increase the safety of socio-technical systems with machine learning components. Four categories of approaches have been identified for promoting safety in general [10]: inherently safe design, safety reserves, safe fail, and procedural safeguards. We find and discuss examples of all of these approaches specifically for machine learning algorithms and especially to mitigate epistemic uncertainty. Through this contribution, we can recommend strategies to engineer safer machine learning methods and set an agenda for further machine learning safety research.

The remainder of the paper is organized in the following manner. In Section II, we discuss harm, risk, uncertainty and the definition of safety. In Section III, we examine statistical machine learning from the safety perspective. Section IV discusses ways to understand the magnitude and direction of harms and sets forth two types of machine learning applications distinguished by their relationship to safety. Section V details a few example applications. Section VI describes ways of achieving safety in general and their specializations to machine learning. Section VII concludes the paper.

II. Definition of Safety

The term *safety* can have many different technical and nontechnical meanings, but for our purposes, we would like to work with a precise, domain-agnostic definition. As well-described in [3], [10] and numerous references therein, such a definition of safety begins with outcomes and events. A system yields an outcome based on its state and the inputs it receives; the outcome event may be desired or undesired. Single events and sets of events have associated costs that can be measured and quantified by society (more on this in Section IV-A). A numeric level of morbidity, for example, can be the cost of an outcome. An undesired outcome is only a harm if its cost exceeds some threshold. Unwanted events of small severity are not counted as safety issues.

The next step in defining safety is to bring in decision theory and the concepts of risk and epistemic uncertainty. Risk is the expected value of the cost of harm: we do not know what the outcome will be, but its distribution is known and we can calculate the expectation of its cost. With uncertainty, we still do not know what the outcome will be, but in contrast to risk, its probability distribution is also unknown (or only partially known). Epistemic uncertainty, in contrast to aleatoric uncertainty, results from lack of knowledge that could be obtained in principle, but may be practically intractable to gather. Some decision theorists argue that all uncertainty can be captured probabilistically, but we maintain the distinction between risk and uncertainty herein, following [3].

Safety is the reduction or minimization of risk and uncertainty of harmful events.

Much can be and is written on costs, risk and uncertainty, and more mathematical precision given. For our purposes, the key points in the definition of safety are that: costs have to be sufficiently high in some human sense for events to be harmful, and that safety involves reducing both the probability of expected harms and the possibility of unexpected harms.

III. Safety and Machine Learning

The starting point in the theory and practice of statistical machine learning is risk minimization. Given joint random variables $X \in \mathcal{X}$ (features) and $Y \in \mathcal{Y}$ (labels) with probability density function $f_{X,Y}(x,y)$, a function mapping $h \in \mathcal{H} : \mathcal{X} \to \mathcal{Y}$, and a loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the risk $R(h)$ is the expectation $\mathbb{E}[L(h(X),Y)] = \int_{\mathcal{X}} \int_{\mathcal{Y}} L(h(x),y) f_{X,Y}(x,y) dy dx$. The loss function L typically measures the discrepancy between the value predicted for y using $h(x)$ and y itself, for example $(h(x)-y)^2$ in regression problems. We would like to find the function h that minimizes the risk.

However, in the machine learning context, we do not have access to the probability $f_{X,Y}$, but rather to a training set of samples drawn i.i.d. from the joint distribution of X and Y: $\{(x_1,y_1),\ldots,(x_m,y_m)\}$. The empirical risk $R_m^{emp}(h)$ is $\frac{1}{m}\sum_{i=1}^{m} L(h(x_i),y_i)$. The empirical risk minimization principle formulates the learning of h as the minimization of $R_m^{emp}(h)$ [5]. Appealing to the results of Glivenko and Cantelli in empirical process theory, it can be shown that the empirical risk $R_m^{emp}(h)$ converges to the risk $R(h)$ uniformly for all h as m goes to infinity. When m is small (in comparison to a suitably defined complexity measure on $\mathcal{H}$), minimizing $R_m^{emp}(h)$ may not yield an h that has small $R(h)$. The structural risk minimization principle alleviates this problem by restricting the complexity of $\mathcal{H}$ or introducing regularization in the minimization problem for h based on some inductive bias.

The risk minimization approach to machine learning has many strengths, as evidenced by the innumerable applied successes it has brought, and captures the risk component of safety. However, it does not capture issues related to uncertainty and loss functions that are relevant for safety. First, although it is assumed that the training samples $\{(x_1,y_1),\ldots,(x_m,y_m)\}$ are drawn from the true underlying probability distribution of (X,Y), that may not always be the case. Furthermore, it may be that the distribution the samples actually come from cannot be known, precluding the use of covariate shift and domain adaptation techniques. This is one form of epistemic uncertainty that is quite relevant to safety because training on a data set from a different distribution can cause much harm.

Also, it may be that the training samples do come from the true, but unknown, underlying distribution, but are absent from large parts of the $\mathcal{X} \times \mathcal{Y}$ space due to small probability density there. Here the learned h will be completely dependent on the inductive bias rather than the uncertain true distribution, which could introduce a safety hazard.

As mentioned above, statistical learning theory analysis utilizes laws of large numbers to study the effect of finite training data and the convergence of the empirical risk to the true risk, but in considering safety, we should also be cognizant that in deployment, a machine learning system only encounters a finite number of test samples and the actual operational risk is an empirical quantity on the test set. Thus the operational risk may be much larger than the true risk for small cardinality test sets, even if h is risk-optimal. This uncertainty caused by the instantiation of the test set can have large safety implications on individual test samples.

As we discussed above, the domain of the loss function in risk minimization is $\mathcal{Y} \times \mathcal{Y}$ and the output is an abstract quantity representing prediction error. However, in real-world applications, the value of the loss function may be endowed with some human cost and that human cost may imply a loss function that also includes $\mathcal{X}$ in the domain. Moreover, the cost may be severe enough to be harmful and thus a safety issue in some parts of the domain and not in others.

IV. Safety Analysis of Machine Learning Applications

As mentioned in Section I, the ways that machine learning enters larger socio-technical systems is myriad. Cyber-physical system applications include robot surgery, self-driving cars, and the smart grid. Decision sciences applications include prison parole, medical treatment, and loan approval. Data products applications include web advertising placement, media recommendation, and spam filtering. Let us see how our definition of machine learning safety applies to these various settings.

A. Harmful Costs

Machine learning safety requires us to first examine an application on whether immediate human costs of its outcomes exceed some severity threshold to be harmful. In decision sciences applications such as the ones listed above, undesired outcomes are truly harmful in a human sense and their effect is felt in near-real time. They are safety issues. Moreover, the space of outcomes is often binary or of small cardinality and it is often self-evident which outcomes are undesired, e.g. prescribing the incorrect medical treatment. However, loss functions are not always monotonic in the correctness of predictions and depend on whose perspective is in the objective. Consider the loan approval application: the applicant would like an approval decision regardless of their features indicating ability to repay, the lender would like approval only in cases in which applicant features indicate likely repayment, and society would like there to be fairness or equitability in the system so that protected groups, such as defined by race and gender, are not discriminated against. The lender perspective is consistent with the typical choice of loss function, but the others are not.

The space of outcomes for the machine learning components of typical cyber-physical systems applications is so vast that it is near-impossible to enumerate all of the outcomes, let alone elicit costs for them. Nevertheless, it is clear that outcomes leading to road accidents, surgical accidents, etc. have high human cost in real time and require the consideration of safety. In order to get more nuanced characterizations of the cost severity of outcomes, a data-driven approach is prudent. As one example in the medical devices domain, we have mined a large database of adverse events to obtain exactly such characterizations [11]. With self-driving cars, such a data-driven approach could help resolve trolley problem-like conundra.

The quality of service implications of unwanted outcomes in data products applications are not typically safety hazards because they do not have an immediate severe human cost. One may argue that an algorithm showing biased or misguided advertisements or a spam filter not allowing an important email to pass could eventually lead to harm, e.g., by being shown an ad for a lower-paying job rather than a higher-paying one, a person may hypothetically end up with a lower quality of life at some point in the future. However, we do not view such a delayed and only hypothetical consequence as a safety issue.

Based only on considering the severity or magnitude of costs of unwanted outcomes, we may divide applications of machine learning into two types: (A) applications in which machine learning algorithms are used to support control and decision making in safety-critical settings with harmful impact on people's lives such a loss of life or injury, and (B) applications in which machine learning based predictions are used in applications with less critical impact. This Type A and Type B nomenclature follows [9]. Cyber-physical systems and decision sciences applications tend to be Type A applications and data products applications tend to be Type B. Table I summarizes the characteristics of Type A and Type B applications along with specific examples from cyber-physical systems, decision sciences, and data products.

B. Epistemic Uncertainty

Safety in machine learning applications can be further analyzed with respect to epistemic uncertainty. There is no a priori reason for the applications to follow the same type structure when examining uncertainty, but as we discuss in the following, the two types are recapitulated.

In addition to the lack of severity of costs, another characteristic of Type B applications is that they are performed at scales with large training sets, large testing sets, and the ability to explore the feature space. For example, in the web advertising

TABLE I
Type A and Type B Applications: Characteristics and Examples

	Type A Applications	Type B Applications
Characteristics	harmful consequences (e.g., death or injury)	less costly consequences (e.g., mission, financial, or information losses)
	many sources of epistemic uncertainties	large scale training and testing sets and the ability to explore the feature space
	real time or near term impact	hypothetical long term impact
Examples	medical robots, autonomous cars, power grids, loan approval, prison sentencing and parole	web advertisement placement, media recommendation, spam filtering

application, one can use billions of data points as training, perform large-scale A/B testing, and evaluate average performance on millions or billions of clicks. For these reasons, the epistemic uncertainties discussed in Section III are less prevalent in Type B applications than in Type A applications. In contrast, in Type A applications it is more often than not the case that there is uncertainty about the training samples being representative of the testing samples, and that only a few predictions are made. Moreover, in cyber-physical systems applications, very large outcome spaces prevent even mild coverage of the space through training samples. Uncertainty of the various types discussed is common in Type A applications.

Thus, not only are errors in Type B applications less costly in human terms, but the amount of uncertainty in the system is less. Therefore, for both reasons, safety is much less relevant in Type B applications than Type A applications. The focus in Type B applications can be squarely on risk minimization whereas Type A applications require the consideration of strategies for achieving safety, as we discuss in Section VI.

V. Detailed Examples

In this section, we further discuss the challenges in ensuring safety in machine learning systems by providing examples from emerging application areas of cyber-physical systems and decision sciences.

A. Cyber-Physical Systems

With advances in computing, networking, and sensing technologies, cyber-physical systems have been deployed in various safety-critical settings such as aerospace, energy, transportation, and healthcare. The increasing complexity and connectivity of these systems, the tight coupling between their cyber- and physical- components, and the inevitable involvement of human operators in their supervision and control has introduced significant challenges in ensuring system reliability and safety while maintaining the expected performance. Cyber-physical systems continuously interact with the physical world and human operators in real-time. In order to adapt to the constantly changing and uncertain environment, they need to take into account not only the current application but also the operator's preferences, intent, and historical behavior [12].

Autonomous machine learning and artificial intelligence techniques have been applied to several decision making and control problems in cyber-physical systems. Here we discuss two examples of Type A applications, where unexpected harmful events with epistemic uncertainty might impact the human lives in real-time.

1) Robot Surgery: Robotically-assisted surgical systems used in minimally invasive surgery are a typical example of human-in-the-loop cyber-physical systems. Surgical robots consist of a teleoperation console operated by a surgeon, an embedded system hosting the robot automated control, and the physical robotic actuators and sensors. The robot control system receives the surgeon's commands issued using the teleoperation console and translates the surgeon's hand, wrist, and finger movements into precisely engineered movements of miniaturized surgical instruments inside patient's body. Recent research shows an increasing interest in use of machine learning algorithms for modeling surgical skills, workflow, and environment and integration of this knowledge into control and automation of surgical robots [13]. Machine learning techniques have been used for detection and classification of surgical motions for automated surgical skill evaluation [14]–[16] and automating portions of surgical tasks (e.g., knot-tying, suturing, or stitching) to reduce the cognitive workload on the surgical team [16], [17].

Although the state-of-the-art surgical robots are human supervised systems that do not take any autonomous actions, there have been ongoing reports on safety incidents during use of such systems that negatively impact patients. Between 2011 and 2013, over 10,000 robotic-surgery-related adverse events were reported to the U.S. Food and Drug Administration (FDA), where majority of these incidents were related to malfunctions in the robot control system and instruments [11]. As surgical robots get enhanced with machine learning algorithms providing real-time technical decision making and autonomous control in the operating room, safety becomes even a bigger concern. Given the uncertainty and large variability in the operator actions and behavior, organs/tissues dynamics, and patient medical history, there are practical limitations in learning surgical trajectories and workflows. The training data often consists of samples collected from a select set of surgical tasks (e.g., elementary suturing gestures) performed by well-trained surgeons, which might not represent the variety of actions and tasks performed during a real procedure. Previous work shows that surgeon's expertise level, surgery type, and medical history have a significant impact on the possibility of complications and errors occurring during surgery.

One solution for dealing with these uncertainties is to assess the robustness of the system in presence of unwanted hazardous events (e.g., failures in control system, noisy sensor measurements, or incorrect commands sent by novice operators) by simulating such events in virtual environments [18] and quantifying the possibility of making safe decisions by the learning algorithm. This assessment would also help with deciding the best safety strategies to be used in design and further refinement of system (see Section VI).

2) Self-Driving Cars: Another example is self-driving cars, which are autonomous cyber-physical systems capable of making intelligent navigation decisions in real-time without any human input. They combine a range of sensor data from laser range finders and radars with the video and GPS data to generate a detailed 3D map of the environment and estimate their position. The control system of the car uses this information to determine the optimal path to destination and sends the relevant commands to actuators that control the steering, braking and throttle. Machine learning algorithms are used in the control system of self-driving cars to model, identify, and track the dynamic environment, including the road conditions and moving objects (e.g., other cars and pedestrians).

Although automated driving systems are expected to eliminate human driver errors and reduce the possibility of crashes, there are several sources of uncertainty and failure that might lead to potential safety hazards in these systems. Unreliable or noisy sensor signals (e.g., GPS data or video signals in bad weather conditions), limitations of computer vision systems, and unexpected changes in the environment (e.g., unknown driving scenes or unexpected accidents on the road) can adversely affect the ability of control system in learning and understanding the environment and making safe decisions [19]. For example, this year a self-driving car (in auto-pilot mode) failed in applying brakes and had a collision with a truck, leading to the death of its driver. This was the first known death event in over 130 million miles of testing the automated driving system. The accident was caused under extremely rare circumstances of the high height of the truck, its white color under the bright sky, combined with the positioning of the cars across the road [20].

The importance of epistemic uncertainty or ”Uncertainty on Uncertainty” in these AI-assisted systems has been recently recognized and there are ongoing research efforts towards quantifying the robustness of self-driving cars to events that are rare (e.g., distance to a bicycle running on an expected trajectory) or not present in the training data (e.g., unexpected trajectories of moving objects) [21].

B. Decision Sciences

In decision sciences applications, people are in the loop in a different way than in cyber-physical systems, but in the loop nonetheless. Decisions are made about people and are made by people using machine learning-based tools for support. Many emerging application domains are now shifting to data-driven decision making due to greater capture of information digitally and the desire to be more scientific rather than relying on (fallible) gut instinct [22]. These applications present many safety-related challenges.

1) Predicting Voluntary Resignation: We recently studied a Type A problem of predicting which IBM employees will voluntarily resign from the company in the next six months based on human resources and compensation data, which required us to develop a classification algorithm to be placed within a larger decision-making system involving human decision makers [23]. There are several sources of epistemic uncertainty in this problem. First, the way to construct a training set in the problem is to look at the historical set of employees and treat employees that voluntarily resigned as positive samples and employees still in the workforce as negative samples. However, since the prediction problem is to predict resignation in the next six months, our set of negative samples will necessarily include employees who should be labeled positively because they will be resigning soon [24]. Another uncertainty is related to quirks or vagaries in the data that are predictive but will not generalize. In this problem, a few predictive features related to stipulations in employees' contracts to remain with IBM for a fixed duration after their company was acquired, but such a pattern would not remain true going forward. Another issue is unique feature vectors: if the data contains an employee in Australia who has gone 17 years without being promoted and no other similar employees, then there is huge uncertainty in that part of feature space and inductive bias must be completely relied upon.

2) Loan Approval: As another example in the decision sciences that we have studied, let us consider the decision to approve loans for solar panels given to the rural poor in India based on data in application forms [25]. Many of the same uncertainties as in the previous example related to the training set being representative of the true test distribution repeat here. An interesting additional issue in this case relates to the human cost function including $\mathcal{X}$. One of the attributes available in the problem was the surname of the applicant. In this part of India, the surname is a strong indicator of religion and caste; for this reason, a hazardous situation may have occurred if surname were used as a feature in the prediction task, even if it helped accuracy.

VI. Strategies for Achieving Safety

As discussed in the introduction, safety is usually investigated on an application-by-application basis and strategies for achieving it the same. For example, setting the minimum thickness of vessels and removing flammable materials from a chemical plant are ways of achieving safety. Analyzing such strategies across domains, [10] has identified four main categories of approaches to achieve safety. In this section, we discuss each of these categories in turn along with specific approaches that extend machine learning formulations beyond risk minimization for safety. In doing so, we must be aware of the timescale under which different applications operate and impact people: not all strategies are applicable in real-time applications such as ones in cyber-physical systems.

A. Inherently Safe Design

Inherently safe design is the exclusion of a potential hazard from the system (instead of controlling the hazard). For example, excluding hydrogen from the buoyant material of a dirigible airship makes it safe. (Another possible safety measure would be to introduce apparatus to prevent the hydrogen from igniting.)

In the machine learning context, we would like robustness against the uncertainty of the training set not being sampled from the test distribution. The training set may have various quirks or biases that are unknown to the user and that will not be present during the test phase. Highly complex modeling techniques used today, including extreme gradient boosting and deep neural networks, may pick up on those data vagaries in the learned models they produce to achieve high accuracy, but might fail due to an unknown shift in the data domain [26].

The models are so complex that it is very difficult to understand how they will react to such shifts and whether they will produce harmful outcomes as a result. Two related ways to introduce inherently safe design is by insisting on models that can be interpreted by people and by excluding features that are not causally-related to the outcome [27]–[30]. By examining interpretable models, features or functions capturing quirks in the data can be noted and excluded, thereby avoiding related harm. Similarly, by excluding non-causal variables, phenomena that are not a part of the true 'physics' of the system can be excluded and related harm avoided. We note that post hoc interpretation of complicated uninterpretable models, appealing for other reasons, does not assure safety via inherently safe design.

The desire for neither interpretability nor causality of models is captured in the standard risk minimization formulation of machine learning. Extra regularization or constraints on $\mathcal{H}$ beyond those implied by structural risk minimization are needed to learn such models. There may be performance loss in accuracy by doing so when measuring accuracy with a common training and testing data probability distribution, but the reduction in epistemic uncertainty by doing so increases safety. Both interpretability and causality may be incorporated into a single learned model, e.g. [31], and causality may be used to induce interpretability, e.g. [32]. In cyber-physical applications with very large outcome spaces such as those employing reinforcement learning, appropriate aggregation of states in outcome policies can make the machine learning interpretable [33].

B. Safety Reserves

A second strategy for achieving safety is through multiplicative or additive reserves, known as safety factors and safety margins, respectively. In mechanical systems, a safety factor is a ratio between the maximal load that does not lead to failure and the load for which the system was designed. Similarly the safety margin is the difference between the two.

For the purposes of machine learning with uncertainty, whether that uncertainty is in the training data matching the test distribution or in the instantiation of the test set, we can parameterize the unknown with the symbol θ. Let the risk of the risk-optimal model if the θ were known be $R^*(\theta)$. Along the same lines as safety factors and safety margins, robust formulations find h while constraining or minimizing $\max_\theta \frac{R(h,\theta)}{R^*(\theta)}$ or $\max_\theta (R(h,\theta) - R^*(\theta))$. Such formulations can capture uncertainty in the class priors and uncertainty resulting from label noise in classification problems [34], [35]. They can also capture the uncertainty of which part of the $\mathcal{X}$ space the actual small set of test samples comes from: we do not care as much about average test error for medical diagnosis problems or self-driving cars if a model will only be used on a handful of patients or road conditions as we do about the maximum test error.

A different sort of safety factor comes about when considering fairness and equitability. In certain prediction problems, the risk of harm for members of protected groups should not be much worse (up to a multiplicative factor) than the risk of harm for others [36]–[38]. Features indicating a protected group, such as race and gender, are dimensions in the $\mathcal{X}$ space; we can partition the space into the sets $\mathcal{X}_p, \mathcal{X}_u \subset \mathcal{X}$ corresponding to the protected and unprotected groups respectively. The safety factor known as disparate impact constrains the following to a minimum value such as $4/5$:

$$\frac{\int_{\mathcal{X}_p} \int_{\mathcal{Y}} L(x, h(x), y) f_{X,Y}(x, y) dy dx}{\int_{\mathcal{X}_u} \int_{\mathcal{Y}} L(x, h(x), y) f_{X,Y}(x, y) dy dx}.$$

Under such a constraint, the risk of harm for protected groups is not much more than for unprotected groups.

C. Safe Fail

The third general category of safety measures is 'safe fail,' which implies that a system remains safe when it fails in its intended operation. Examples are electrical fuses, so-called dead man's switches on trains, and safety valves on boilers.

A technique used in machine learning when predictions cannot be given confidently is the reject option [39]: the model reports that it cannot reliably give a prediction and does not attempt to do so, thereby failing safely. When the model elects the reject option, typically a human operator intervenes, examines the test sample, and provides a manual prediction.

In classification problems, models are reported to be least confident near the decision boundary. However, by doing so, there is an implicit assumption that distance from the decision boundary is inversely related to confidence. This is reasonable in parts of $\mathcal{X}$ with high probability density and large numbers of training samples because the decision boundary is located where there is a large overlap in likelihood functions. However, as discussed in Section III, parts of $\mathcal{X}$ with low density may not contain any training samples at all and the decision boundary may be completely based on an inductive bias, thereby containing much epistemic uncertainty. In these parts of the space, distance from the decision boundary is fairly meaningless and the typical trigger for the reject option should be avoided [40]. For a rare combination of features in a test sample [41], a safe fail mechanism is to always go for manual examination.

Both of these manual intervention options are applicable to decision sciences applications in which the timescale is longer than in cyber-physical systems. When working on the scale of milliseconds, only options similar to dead man's switches that stop operations in a reasonable manner are applicable.

D. Procedural Safeguards

Finally, the fourth strategy for achieving safety is given the name procedural safeguards. This strategy includes measures beyond ones designed into the core functionality of the system, such as audits, training, posted warnings, and so on. Two directions in machine learning that can be used for increasing safety within this category are user experience design and openness.

In Type A decision sciences applications especially, non-specialists are often the operators of machine learning systems. Defining the training data set and setting up evaluation procedures, among other things, have certain subtleties that can cause harm during operation if done incorrectly. User experience design can be used to guide and warn novice and experienced practitioners to set up machine learning systems properly and thereby increase safety.

Best of breed machine learning algorithms these days are open source, which allows for the possibility of public audit. Safety hazards and potential harms can be discovered through examination of source code. However, open source software is not enough, because the behavior of machine learning systems is driven by data as much as it is by software implementations of algorithms. Open data refers to data that can be freely used, reused and redistributed by anyone. It is more common in Type A applications such as those sponsored or run by governments than in Type B applications where the data is oftentimes the key value proposition. Opening data is a procedural safeguard for increasing safety that is increasingly being adopted in Type A applications [42]–[44].

VII. Conclusion

Machine learning systems are already embedded in many functions of society. The prognosis is for broad adoption to only increase across all areas of life. With this prevailing trend, machine learning researchers, engineers, and ethicists have started discussing the topic of safety. In this paper, we contribute to this discussion starting from a very basic definition of safety in terms of harm, risk, and uncertainty and building upon it in the machine learning context. We identify that the minimization of epistemic uncertainty is missing from standard modes of machine learning developed around risk minimization and that it needs to be included when considering safety. We have delineated two types of applications of machine learning: Type A in which safety is an important concern and Type B in which it is not so. We have discussed several strategies for increasing safety that are especially pertinent in Type A applications.

The strategies for increasing safety that we mentioned in Section VI are not a comprehensive list and are far from fully developed. This paper can be seen as laying the foundations for a research agenda motivated by Type A applications and safety within which further strategies can be developed and existing strategies can be fleshed out. In some respects, the research community has taken risk minimization close to the limits of what is achievable. Safety, especially epistemic uncertainty minimization, represents a direction that offers new and exciting problems to pursue. As it is said in the Sanskrit literature, *ahiṃsā paramo dharmaḥ* (non-harm is the ultimate direction). Moreover, not only is non-harm the first ethical duty, many of the safety issues for machine learning we have discussed in this paper are starting to enter legal obligations as well [45].

References

[1] A. Conn, "The AI wars: The battle of the human minds to keep artificial intelligence safe," http://futureoflife.org/2015/12/17/the-ai-wars-the-battle-of-the-human-minds-to-keep-artificial-intelligence-safe, Dec. 2015.

[2] T. Ferrell, "Engineering safety-critical systems in the 21st century," 2010.

[3] N. Möller, "The concepts of risk and safety," in *Handbook of Risk Theory*, S. Roeser, R. Hillerbrand, P. Sandin, and M. Peterson, Eds. Dordrecht, Netherlands: Springer, 2012, pp. 55–85.

[4] K. R. Varshney, "Engineering safety in machine learning," in *Proc. Inf. Theory Appl. Workshop*, La Jolla, CA, Feb. 2016.

[5] V. Vapnik, "Principles of risk minimization for learning theory," in *Adv. Neur. Inf. Process. Syst. 4*, 1992, pp. 831–838.

[6] K. L. Wagstaff, "Machine learning that matters," in *Proc. Int. Conf. Mach. Learn.*, Edinburgh, United Kingdom, Jun.–Jul. 2012, pp. 529–536.

[7] H. Alemzadeh, "Data-driven resiliency assessment of medical cyber-physical systems," Ph.D. dissertation, Univ. Illinois, Urbana-Champaign, Urbana, IL, 2016.

[8] J. Stanley and D. Tunkelang, "Doing data science right — your most common questions answered," http://firstround.com/review/doing-data-science-right-your-most-common-questions-answered, 2016.

[9] K. R. Varshney, "Data science of the people, for the people, by the people: A viewpoint on an emerging dichotomy," in *Proc. Data for Good Exchange Conf.*, New York, NY, Sep. 2015.

[10] N. Möller and S. O. Hansson, "Principles of engineering safety: Risk and uncertainty reduction," *Reliab. Eng. Syst. Safe.*, vol. 93, no. 6, pp. 798–805, Jun. 2008.

[11] H. Alemzadeh, J. Raman, N. Leveson, Z. Kalbarczyk, and R. K. Iyer, "Adverse events in robotic surgery: A retrospective study of 14 years of FDA data," *PLoS ONE*, vol. 11, no. 4, pp. 1–20, 04 2016.

[12] G. Schirner, D. Erdogmus, K. Chowdhury, and T. Padir, "The future of human-in-the-loop cyber-physical systems," *Computer*, no. 1, pp. 36–45, 2013.

[13] Y. Kassahun, B. Yu, A. T. Tibebu, D. Stoyanov, S. Giannarou, J. H. Metzen, and E. Vander Poorten, "Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions," *International Journal of Computer Assisted Radiology and Surgery*, vol. 11, no. 4, pp. 553–568, 2016.

[14] H. C. Lin, I. Shafran, T. E. Murphy, A. M. Okamura, D. D. Yuh, and G. D. Hager, *Automatic Detection and Segmentation of Robot-Assisted Surgical Motions.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 802–810.

[15] H. C. Lin, I. Shafran, D. Yuh, and G. D. Hager, "Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions," *Computer Aided Surgery*, vol. 11, no. 5, pp. 220–230, 2006.

[16] C. E. Reiley, E. Plaku, and G. D. Hager, "Motion generation of robotic surgical tasks: Learning from expert demonstrations," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, Aug 2010, pp. 967–970.

[17] A. Shademan, R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger, and P. C. W. Kim, "Supervised autonomous robotic soft tissue surgery," *Science Translational Medicine*, vol. 8, no. 337, pp. 37ra64–337ra64, 2016.

[18] H. Alemzadeh, D. Chen, A. Lewis, Z. Kalbarczyk, J. Raman, N. Leveson, and R. Iyer, "Systems-theoretic safety assessment of robotic telesurgical systems," in *International Conference on Computer Safety, Reliability, and Security*. Springer, 2015, pp. 213–227.

[19] S. Rayej, "How do self-driving cars work?" http://robohub.org/how-do-self-driving-cars-work/, 2014.

[20] J. Lowy, "Driver killed in self-driving car accident for first time," http://www.pbs.org/newshour/rundown/driver-killed-in-self-driving-car-accident-for-first-time/, 2016.

[21] J. Duchi, P. Glynn, and R. Johari, "Uncertainty on uncertainty, robustness, and simulation," http://aicenter.stanford.edu/uncertainty-on-uncertainty-robustness-and-simulation/, SAIL-Toyota Center for AI Research at Stanford.

[22] E. Brynjolfsson, L. Hitt, and H. Kim, "Strength in numbers: How does data-driven decision-making affect firm performance?" in *Proc. Int. Conf. Inf. Syst.*, Shanghai, China, Dec. 2011, p. 13.

[23] M. Singh, K. R. Varshney, J. Wang, A. Mojsilović, A. R. Gill, P. I. Faur, and R. Ezry, "An analytics approach for proactively combating voluntary attrition of employees," in *Proc. IEEE Int. Conf. Data Min. Workshops*, Brussels, Belgium, Dec. 2012, pp. 317–323.

[24] D. Wei and K. R. Varshney, "Robust binary hypothesis testing under contaminated likelihoods," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Brisbane, Australia, Apr. 2015, pp. 3407–3411.

[25] H. Gerard, K. Rao, M. Simithraaratchy, K. R. Varshney, K. Kabra, and G. P. Needham, "Predictive modeling of customer repayment for sustainable pay-as-you-go solar power in rural India," in *Proc. Data for Good Exchange Conf.*, New York, NY, Sep. 2015.

[26] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proc. ACM SIGKDD Conf. Knowl. Discov. Data Min.*, Sydney, Australia, Aug. 2015, pp. 1721–1730.

[27] A. A. Freitas, "Comprehensible classification models – a position paper," *SIGKDD Explorations*, vol. 15, no. 1, pp. 1–10, Jun. 2013.

[28] C. Rudin, "Algorithms for interpretable machine learning," in *Proc. ACM SIGKDD Conf. Knowl. Discov. Data Min.*, New York, NY, Aug. 2014, p. 1519.

[29] S. Athey and G. W. Imbens, "Machine learning methods for estimating heterogeneous causal effects," http://arxiv.org/pdf/1504.01132.pdf, Jul. 2015.

[30] M. Welling, "Are ML and statistics complementary?" in *IMS-ISBA Meeting on 'Data Science in the Next 50 Years'*, Dec. 2015.

[31] F. Wang and C. Rudin, "Causal falling rule lists," http://arxiv.org/pdf/1510.05189.pdf, Oct. 2015.

[32] A. Chakarov, A. Nori, S. Rajamani, S. Sen, and D. Vijaykeerthy, "Debugging machine learning tasks," http://arxiv.org/pdf/1603.07292.pdf, Mar. 2016.

[33] M. Petrik and R. Luss, "Interpretable policies for dynamic product recommendations," in *Proc. Conf. Uncertainty Artif. Intell.*, Jersey City, NJ, Jun. 2016, p. 74.

[34] F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Mach. Learn.*, vol. 42, no. 3, pp. 203–231, Mar. 2001.

[35] M. A. Davenport, R. G. Baraniuk, and C. D. Scott, "Tuning support vector machines for minimax and Neyman-Pearson classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1888–1898, Oct. 2010.

[36] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 7, pp. 1445–1459, Jul. 2013.

[37] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proc. ACM SIGKDD Conf. Knowl. Discov. Data Min.*, Sydney, Australia, Aug. 2015, pp. 259–268.

[38] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Rev.*, vol. 104, 2016.

[39] K. R. Varshney, R. J. Prenger, T. L. Marlatt, B. Y. Chen, and W. G. Hanley, "Practical ensemble classification error bounds for different operating points," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 11, pp. 2590–2601, Nov. 2013.

[40] J. Attenberg, P. Ipeirotis, and F. Provost, "Beat the machine: Challenging humans to find a predictive model's "unknown unknowns"," *ACM J. Data Inf. Qual.*, vol. 6, no. 1, p. 1, Mar. 2015.

[41] G. M. Weiss, "Mining with rarity: A unifying framework," *SIGKDD Explor. Newsletter*, vol. 6, no. 1, pp. 7–19, Jun. 2004.

[42] A. Sahuguet, J. Krauss, L. Palacios, and D. Sangokoya, "Open civic data: Of the people, by the people, for the people," *Bull. Tech. Comm. Data Eng.*, vol. 37, no. 4, pp. 15–26, Dec. 2014.

[43] E. Shaw, "Improving service and communication with open data: A history and how-to," Ash Center, Harvard Kennedy School, Tech. Rep., Jun. 2015.

[44] S. Kapoor, A. Mojsilović, J. N. Strattner, and K. R. Varshney, "From open data ecosystems to systems of innovation: A journey to realize the promise of open data," in *Proc. Data for Good Exchange Conf.*, New York, NY, Sep. 2015.

[45] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation'," in *Proc. ICML Workshop Human Interpretability*, New York, NY, Jun. 2016, pp. 26–30.