Big Data
Volume 5 Number 3, 2017

DOI: 10.1089/big.2016.0051

ORIGINAL ARTICLE

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

Kush R. Varshney[1,*] and Homa Alemzadeh[2]

Abstract

Machine learning algorithms increasingly influence our decisions and interact with us in all parts of our daily lives. Therefore, just as we consider the safety of power plants, highways, and a variety of other engineered socio-technical systems, we must also take into account the safety of systems involving machine learning. Heretofore, the definition of safety has not been formalized in a machine learning context. In this article, 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. We find that the foundational principle of modern statistical machine learning, empirical risk minimization, is not always a sufficient objective. We discuss how four different categories of strategies for achieving safety in engineering, including inherently safe design, safety reserves, safe fail, and procedural safeguards can be mapped to a machine learning context. We then discuss example techniques that can be adopted in each category, such as considering interpretability and causality of predictive models, objective functions beyond expected prediction accuracy, human involvement for labeling difficult or rare examples, and user experience design of software and open data.

Keywords: cyber-physical systems; data products; decision science; machine learning; safety

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. However, 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] Safety is a notion that is domain specific, cf. safe food and water, safe vehicles and highways, safe medical treatments, safe toys, safe neighborhoods, and safe industrial plants. Each of these domains has specific design principles and regulations that are applicable only to them.

There are some loose notions of safety for machine learning, but they are primarily of the "I know it when I see it" variety or are very application specific; to the best of our knowledge,[3] there is no precise, nonapplication-specific, first-principles definition of safety for machine learning. The main contribution of this article is to provide exactly such a definition. To do so, we build upon a universal domain-agnostic definition of safety in the engineering literature.[4,5]

In Refs.,[4,5] and numerous references therein, Möller et al. propose a decision-theoretic definition of safety

[1]*Department of Data Science, IBM Thomas J. Watson Research Center, Yorktown Heights, New York.*
[2]*Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, Virginia.*

*Address correspondence to: *Kush R. Varshney, IBM Thomas J. Watson Research Center, 1101 Kitchawan Rd, Yorktown Heights, NY 10598,* E-mail: krvarshn@us.ibm.com

that applies to a broad set of domains and systems. They define safety to be the reduction or minimization of *risk* and *epistemic uncertainty* associated with unwanted outcomes that are severe enough to be seen as *harmful*. The key points in this definition are (1) the cost of unwanted outcomes has to be sufficiently high in some human sense for events to be harmful and (2) safety involves reducing both the probability of expected harms and the possibility of unexpected harms.

We define safety in machine learning in the same way, as the minimization of both risk and uncertainty of harms, and devote the next section to fleshing out the details of this definition. As such, formulations of machine learning for achieving safety that we describe in the Strategies for Achieving Safety section must have both risk and uncertainty minimization in their objective functions explicitly, implicitly through constraints, or through socio-technical components beyond the core machine learning algorithm. The harmful cost regime is the part of the space that requires the dual objectives of risk and uncertainty minimization; the non-harmful cost regime does not require the uncertainty minimization objective.

As background before getting to those sections, we briefly describe harms, risk, and uncertainty without specialization to machine learning. A system yields an outcome based on its state and the inputs it receives. An outcome event may be desired or undesired. Single events and sets of events have associated costs that can be measured and quantified by society. For example, a numeric level of morbidity 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. Risk is the expected value of the cost. Epistemic uncertainty results from the lack of knowledge that could be obtained in principle, but may be practically intractable to gather.[6] Harmful outcomes often occur in regimes and operating conditions that are unexpected or undetermined. With risk, 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). Some decision theorists argue that all uncertainty can be captured probabilistically, but we maintain the distinction between risk and uncertainty.[5]

The first contribution of this work is to critically examine the foundational statistical machine learning principles of empirical risk minimization and structural risk minimization[7] from the perspective of safety. We discuss how they do not deal with epistemic uncertainty. Furthermore, these principles rely on arguments involving average losses and laws of large numbers, which may not necessarily be fully applicable when considering safety. Moreover, the loss functions involved in these principles are abstract measures of distance between true and predicted values rather than application-specific quantities measuring the possibility of outcomes such as loss of life or loss of quality of life that can be judged harmful or not.[8]

A discussion of safety would be incomplete without 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[4]: inherently safe design, safety reserves, safe fail, and procedural safeguards. As a second contribution, we discuss 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 third contribution of this article is examining the definition of and strategies for safety in specific machine learning applications. Today, machine learning technologies are 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, for example, surgical robots, self-driving cars, and the smart grid.[9] By decision sciences, we mean the use of algorithms to aid people in making important decisions and informing strategy, for example, prison parole, medical treatment, and loan approval.[10] By data products, we mean the use of algorithms to automate informational products, for example, web advertising placement, media recommendation, and spam filtering.[10] These settings vary widely in terms of their interaction with people, the scale of data, the time scale of operation and consequence, and the cost magnitude of consequences. 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,[11] we suggest a data-driven approach for characterizing harms that are particularly relevant for cyber-physical systems with large state spaces of outcomes.

Overall, the purpose of this article is to introduce a common language and framework for understanding, evaluating, and designing machine learning systems that involve society and technology. Our goal is to set forth a fundamental organizing and unifying principle that carries through to abstract theoretical formulations of machine learning as well as to concrete real-world applications of machine learning. Thus, it provides practitioners working at any level of abstraction a principled way to reason about the space of socio-technical solutions.

The remainder of the article is organized in the following manner: in the Safety in Machine Learning section, after introducing the standard notation and concept of statistical machine learning, we discuss what harm, risk, and epistemic uncertainty mean for machine learning. In the Strategies for Achieving Safety section, we discuss specific strategies for achieving safety in machine learning. The Example Applications section dives into example applications in cyber-physical systems, decision sciences, and data products. The Conclusion section concludes the article.

Safety in Machine Learning

In this section, after briefly introducing statistical machine learning notation, we examine how machine learning applications fit with the conception of safety given above.

Notation

In what follows, we use standard notation to describe concepts from empirical risk minimization.[7] 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 defined as the expected value of loss:

$$\mathbb{E}[L(h(X), Y)] = \int_{\mathcal{X}} \int_{\mathcal{Y}} L(h(x), y) f_{X,Y}(x, y) dy dx.$$

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 learn the function h that minimizes the risk.

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 (X, Y): $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ and the goal is to learn h such that the empirical risk $R_m^{emp}(h)$ is minimized. The empirical risk is given by:

$$R_m^{emp}(h) = \frac{1}{m} \sum_{i=1}^{m} L(h(x_i), y_i).$$

Harmful costs

Analyzing safety requires us first to examine whether immediate human costs of outcomes exceed some severity threshold to be harmful. Unlike other domains mentioned in the introduction, such as safe industrial plants and safe toys, we have a great advantage when working with machine learning systems because the optimization formulation explicitly includes the loss function L. The domain of L is $\mathcal{Y} \times \mathcal{Y}$ and the output is an abstract quantity representing prediction error. 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.

In many decision science applications, 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. However, loss functions are not always monotonic in the correctness of predictions and depend on whose perspective is in the objective. 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 accidents have high human cost in real time and require the consideration of safety. To get more nuanced characterizations of the cost severity of outcomes, a data-driven approach is prudent.[12] The quality of service implications of unwanted outcomes in data product applications are not typically safety hazards because they do not have an immediate severe human cost. Undesired outcomes may only hypothetically lead to human consequences. In practice, often the acceptable levels of safety and accident rates are defined by the society and the application domain. For example, the difference in acceptable accident rates and costs in motor vehicles (hundreds of thousands of fatalities per year) versus commercial aircraft (tens of fatalities per year) shows the subjectivity of the public's acceptance of safety.[13]

Risk and epistemic uncertainty

The risk minimization approach to machine learning has many strengths, which are evident by its successful application in various domains. We benefit from this explicit optimization formulation in the machine learning domain by automatically reducing the probability of harms, which is not always the case in other domains. However, this standard formulation does not capture the issues related to uncertainty that are also 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 generating the samples cannot be known, precluding the use of covariate shift[14] and domain adaptation techniques.[15] This is one form of epistemic uncertainty that is quite relevant to safety because training on a dataset 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 function h will be completely dependent on an inductive bias encoded through $\mathcal{H}$ rather than the uncertain true distribution, which could introduce a safety hazard.

Statistical learning theory analysis utilizes laws of large numbers to study the effect of finite training data and the convergence of $R_m^{emp}(h)$ to $R(h)$. However, when considering safety, we should also be cognizant that in practice, 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 actual 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.

Applications performed at scales with large training sets, large testing sets, and the ability to explore the feature space have little epistemic uncertainty, whereas in other applications, it is more often than not the case that there is uncertainty about the training samples being representative of the testing samples and only a few predictions are made. Moreover, in applications such as cyber-physical systems, very large outcome spaces prevent even mild coverage of the space through training samples.

Strategies for Achieving Safety

As discussed, safety and strategies for achieving it are often investigated on an application-by-application basis. 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, Ref.[4] has identified four main categories of approaches to achieve safety.

First, 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.)

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.

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.

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.

In this section, we discuss each of these strategies with specific approaches that extend machine learning formulations beyond risk minimization for safety.

Inherently safe design

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 biases that are unknown to the user and will not be present during the test phase, or may contain patterns that are undesired and might lead to harmful outcomes. Modern techniques such as extreme gradient boosting and deep neural networks may exploit these biases and achieve high accuracy, but they may fail in making safe predictions due to unknown shifts in the data domain or inferring incorrect patterns or harmful rules.[16]

These 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 are by insisting on models that can be interpreted by people and excluding features that are not causally related to the outcome.[17–20] By examining interpretable models, features or functions capturing quirks in the data can be noted and excluded, thereby avoiding related harm. Similarly, by carefully selecting variables that are causally related to the outcome, phenomena that are not a part of the true "physics" of the system can be excluded, and associated harm be avoided. We note that *post hoc* interpretation of complex uninterpretable models, appealing for other reasons, does not assure safety by inherently safe design because the interpretation is not the decision rule that is actually used in making predictions.

Neither interpretability nor causality of models is properly captured within 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 inherently safe models. This might lead to performance loss in accuracy when measured with a common training and testing data probability distribution, but the safety will be enhanced by reduction in epistemic uncertainty. Both interpretability and causality may be incorporated into a single learned model,[21] and causality may be used to induce interpretability.[22] In applications with very large outcome spaces such as those employing reinforcement learning, it is shown that appropriate aggregation of states in outcome policies can lead to interpretable models.[23]

Safety reserves

In machine learning formulations, the uncertainty in the matching of training and test data distributions or in the instantiation of the test set can be parameterized with the symbol θ. Let $R^*(\theta)$ be the risk of the risk-optimal model if the θ were known. 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.[24,25] They can also capture the uncertainty of which part of the $\mathcal{X}$ space the actual small set of test samples comes from.

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.[26–28] We can partition the feature space $\mathcal{X}$ into the sets $\mathcal{X}_u, \mathcal{X}_p \subset \mathcal{X}$, respectively, corresponding to the unprotected and protected groups, indicated by features such as race and gender. Then using a rule such as the 80% (or four-fifths) rule advocated in the study of disparate impact,[29] we can constraint the relative risk of harm for the protected versus unprotected group to a maximum value such as 5/4:

$$\frac{\int_{\mathcal{X}_p}\int_{\mathcal{Y}} L(x,h(x),y) f_{X,Y}(x,y)dydx}{\int_{\mathcal{X}_u}\int_{\mathcal{Y}} L(x,h(x),y) f_{X,Y}(x,y)dydx} \leq \frac{5}{4}.$$

Under such a constraint, we ensure that the outcome of prediction for protected groups is not much more harmful than for unprotected groups.

Safe fail

A technique used in machine learning when predictions cannot be given confidently is the reject option[30]: the model reports that it cannot reliably give a prediction and does not attempt to do so, thereby failing safely. When the model selects 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, 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.[31] For a rare combination of features in a test sample,[32] a safe fail mechanism is to always go for manual examination.

Both of these manual intervention options are suitable for applications with sufficiently long time scales. When working on the scale of milliseconds, only options similar to dead man's switches that stop operations in a reasonable manner are applicable.

Procedural safeguards

In addition to general procedural safeguards that carry over from other domains, two directions in machine learning that can be used for increasing safety within this category are user experience design and openness.

In decision science applications especially, nonspecialists 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.

These days, most of the best machine learning algorithms 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 sufficient, because the behavior of machine learning systems is driven by data as much as it is driven by software implementations of algorithms. Open data refer to data that can be freely used, reused, and redistributed by anyone. Opening data is a procedural safeguard for increasing safety that is increasingly being adopted by the community.[33–35]

Example Applications

In this section, we further detail safety in machine learning systems by providing examples from cyber-physical systems, decision sciences, and data products.

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 have 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. 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 past behavior.[36]

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

Surgical robots. Robotically assisted surgical systems 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 automated robot 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 the use of machine learning algorithms for modeling surgical skills, workflow, and environment and integration of this knowledge into control and automation of surgical robots.[37] Machine learning techniques have been used for detection and classification of surgical motions for automated surgical skill evaluation[38–40] and automating portions of repetitive and time-consuming surgical tasks (e.g., knot-tying and suturing).[40,41]

In autonomous robotic surgery, a machine learning-enabled surgical robot continuously estimates the state of the environment (e.g., length or thickness of soft tissues under surgery) based on the measurements from sensors (e.g., image data or force signals) and generates a plan for executing actions (e.g., moving the robotic instruments along a trajectory). The mapping function from the perception of environment to the robotic actions is considered a surgical skill that the robot learns, through evaluation of its own actions or by observing the actions of expert surgeons. The quality of the learned surgical skills can be assessed using cost functions that are either automatically learned or manually defined by surgeons.[37]

Given the uncertainty and large variability in the operator actions and behavior, organ/tissue movements and dynamics, and possibility of incidental failures in the robotic system and instruments, predicting all possible system states and outcomes and assessing their associated costs are very challenging. As mentioned in the Harmful Costs section, due to the very large outcome space, it is not straightforward to elicit costs of all different outcomes and characterize which tasks or actions are costly enough to represent safety issues. For example, there have been ongoing reports of safety

incidents during use of surgical robots that negatively impact patients by causing procedure interruptions or minor injuries. These incidents happen despite existing safe-fail mechanisms included in the system and often result from a combination of different causal factors and unexpected conditions, including malfunctions of surgical instruments, actions taken by the surgeon, and the patient's medical history.[12]

There are also practical limitations in learning optimal and safe surgical trajectories and workflows due to epistemic uncertainty in such environments. The training data often consist 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. Furthermore, automated algorithms should be able to cope with uncertainty and unpredictable events and guarantee patient safety, just as expert surgeons do in such scenarios.[37]

One solution for dealing with these uncertainties is to assess the robustness of the system in the presence of unwanted and rare 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[42] and quantifying the possibility of making safe decisions by the learning algorithm. This approach is an example of procedural safeguards (Procedural Safeguards section). Such a simulated assessment also serves to highlight the situations requiring safe-fail strategies, such as converting the procedure to nonrobotic techniques, rescheduling it to a later time, or restarting the system, which can refine the system.

The costs of unwanted outcomes and safe-fail strategies to cope with them can also be characterized based on past data. For example, we mined the FDA's Manufacturer and User Facility Device Experience (MAUDE) database, a large database containing 14 years worth of adverse events, to obtain such characterizations on the causes and severity of safety incidents and recovery actions taken by the surgical team. Such analysis helps focus on development of machine learning algorithms containing safety strategies on regimes with harmful outcomes and avoid concern for safety strategies in regimes with nonharmful outcomes.

Another solution currently adopted in practice is through supervisory control of automated surgical tasks instead of fully autonomous surgery. For example, if the robot generates a geometrically optimized suture plan based on sensor data or surgeon input, it should still be tracked and updated in real time because of possible tissue motion and deformation during surgery.[41] This is an example of examining interpretable models to avoid possible harm (as discussed in the Inherently Safe Design section). An example of adopting safety reserves (Safety Reserves section) in robotic surgery is robust optimization of preoperative planning to minimize the uncertainty at the task level, while maximizing the dexterity.[43]

Self-driving cars. Self-driving cars 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 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 the 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.[44] For example, a self-driving car (in autopilot mode) recently collided with a truck after failing to apply the brakes, leading to the death of the truck driver. This was the first known fatality 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.[45]

The importance of epistemic uncertainty or "uncertainty on uncertainty" in these AI-assisted systems has been recently recognized, and there are ongoing

research efforts toward 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).[46] Systems that recognize such rare events trigger safe-fail mechanisms.

To the best of our knowledge, there is no self-driving car system with an inherently safe design that utilizes, for example, interpretable models.[47] Fail-safe mechanisms that upon detection of failures or less confident predictions stop the autonomous control software and switch to a backup system or a degraded level of autonomy (e.g., full control by the driver) are considered for self-driving cars.[48]

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 by people using machine learning-based tools for support. Many emerging application domains are now shifting to data-driven decision making due to a greater capture of information digitally and the desire to be more scientific rather than relying on (fallible) gut instinct.[49] These applications present many safety-related challenges.

Predicting voluntary resignation. We recently studied the problem of predicting which IBM employees will voluntarily resign from the company in the next 6 months based on human resource and compensation data, which required us to develop a classification algorithm to be placed within a larger decision-making system involving human decision makers.[50] 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 who 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 6 months, our set of negative samples will necessarily include employees who should be labeled positively because they will be resigning soon.[51]

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.

In the solution created for this problem, the inherently safe design principle of interpretability (Inherently Safe Design section) was insisted upon and was what led to the discovery about the acquired company. Specifically, C5.0 decision trees were used with the rule set option, and the project directly motivated the study of an optimization approach for learning classification rules.[52] The reason for conducting the project was to take actions such as salary increases to retain employees at risk of resigning, and for this, the other inherently safe design principle of causality is important. Rare samples such as the Australian employee led to the safe-fail mechanism of manual inspection.

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.[53] The epistemic uncertainty related to the training set not being representative of the true test distribution repeats here and can be addressed by similar safety strategies as discussed in the previous examples.

Loan approval is an example illustrating loss functions that are not always monotonic in the correctness of predictions and depend on perspective. 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 gender and religion, are not discriminated against. The lender perspective is consistent with the typical choice of the loss function, but the others are not.

An interesting additional issue, in this case, relates to the human cost function from society's perspective, 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. The use of this variable as a feature improved classification accuracy by a couple of percentage points, but resulted in worse fairness: the true cost in the problem from society's perspective. Simply dropping the attribute as a feature does not ensure fairness because

other features may be correlated, but a safety margin on the accuracy of the groups makes the system fairer.

Data products

With data product applications, the first question to consider is whether immediate costs are large enough for them to be considered safety issues. 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. For example, 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. Here the cost function does depend on $\mathcal{X}$ because misclassifying certain e-mails is more costly than others. However, we do not view such a delayed and only hypothetical consequence as a safety issue.

Moreover, in typical data product applications, 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. Therefore, uncertainty is not at the forefront, and neither are the safety strategies. For example, the procedural safeguard of opening data is more common in decision science applications such as those sponsored or run by governments than in data product applications where the data is often the key value proposition.

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, researchers, engineers, and ethicists have started discussing the topic of safety in machine learning. In this article, 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 statistical risk minimization and it needs to be included when considering safety.

We discuss a few strategies for increasing safety in machine learning that are not a comprehensive list and are far from fully developed. This article can be seen as laying the foundations for a research agenda motivated by safety within which further strategies can be developed and existing strategies can be flushed 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, many of which are being pursued already. As it is said in the Sanskrit literature, *ahiṃsā paramo dharmaḥ* (nonharm is the ultimate direction).

Moreover, not only is nonharm the first ethical duty, many of the safety issues for machine learning we have discussed in this article are starting to enter legal obligations as well. For example, the European Union has recently adopted a set of comprehensive regulations for data protection, which include prohibiting algorithms that make any "decision based solely on automated processing, including profiling" and significantly affect a data subject, or produce legal effects concerning him/her. This regulation, which will take effect in 2018, is anticipated to restrict a wide range of machine learning algorithms currently used in, for example, recommendation systems, credit and insurance risk assessments, and social networks.[54]

We present example applications where machine learning algorithms are increasingly used and discuss the aspects of epistemic uncertainty, harmful outcomes, and potential strategies for achieving safety for each application. In some applications such as cyber-physical systems and decision sciences, machine learning algorithms are used to support control and decision making in safety-critical settings with considerable costs and direct harmful impact on people's lives, such as injury or loss of life. In other applications, machine learning-based predictions are only used in less critical settings for automated informational products. Applications with higher costs of unwanted outcomes tend to be also those with higher uncertainty and the ones with less severe outcomes are the ones with smaller uncertainty.

Author Disclosure Statement

No competing financial interests exist.

References

1. Conn A. 2015. The AI wars: The battle of the human minds to keep artificial intelligence safe. Available online at http://futureoflife.org/2015/12/17/the-ai-wars-the-battle-of-the-human-minds-to-keep-artificial-intelligence-safe (accessed September 8, 2017).
2. Ferrell T. Engineering safety-critical systems in the 21st century. IEEE Central Virginia Section, Engineers Week Dinner Meeting. Charlottesville, VA, 2010.
3. Varshney KR. Engineering safety in machine learning. In: Proceeding of Information Theory and Application Workshop, La Jolla, CA, 2016.
4. Möller N, Hansson SO. Principles of engineering safety: Risk and uncertainty reduction. Reliab Eng Syst Safe. 2008;93:798–805.
5. Möller N. The concepts of risk and safety. In: Roeser S, Hillerbrand R, Sandin P, Peterson M (Eds.): Handbook of Risk Theory, Dordrecht, Netherlands: Springer, 2012. pp. 55–85.
6. Senge R, Bösner S, Dembczynski K, et al. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. Inf Sci. 2014;255:16–29.
7. Vapnik V. Principles of risk minimization for learning theory. Adv Neur Inf Process Syst. 1992;4:831–838.

8. Wagstaff KL. Machine learning that matters. In: Proceedings of International Conference on Machine Learning, Edinburgh, United Kingdom, June–July 2012, pp. 529–536.
9. Alemzadeh H. Data-driven resiliency assessment of medical cyber-physical systems. Ph.D. dissertation, University of Illinois, Urbana-Champaign, Urbana, IL, 2016.
10. Stanley J, Tunkelang D. 2016. Doing data science right—your most common questions answered. Available online at http://firstround.com/review/doing-data-science-right-your-most-common-questions-answered (accessed September 8, 2017).
11. Olteanu A, Talamadupula K, Varshney KR. The limits of abstract evaluation metrics: The case of hate speech detection. In: Proceedings on ACM Web Science Conference, Troy, NY, 2017, pp. 405–406.
12. Alemzadeh H, Raman J, Leveson N, et al. Adverse events in robotic surgery: A retrospective study of 14 years of FDA data. PLoS One. 2016;11:1–20.
13. Knight J. Fundamentals of dependable computing for software engineers. Boca Raton, FL: CRC Press, 2012.
14. Shimodaira H. Improving predictive inference under covariate shift by weighting the log-likelihood function. J Stat Plan Inference. 2000;90:227–244.
15. Daume H III, Marcu D. Domain adaptation for statistical classifiers. J Artif Intell Res 2006;26:101–126.
16. Caruana R, Lou Y, Gehrke J, et al. Elhadad, intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings on ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 2015, pp. 1721–1730.
17. Freitas AA. Comprehensible classification models—A position paper. SIGKDD Explorations 2013;15:1–10.
18. Rudin C.. Algorithms for interpretable machine learning. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, 2014, p. 1519.
19. Athey S, Imbens GW. 2015. Machine learning methods for estimating heterogeneous causal effects. Available online at http://arxiv.org/pdf/1504.01132.pdf (accessed September 8, 2017).
20. Welling M. 2015. Are ML and statistics complementary? In: IMS-ISBA Meeting on 'Data Science in the Next 50 Years'.
21. Wang F, Rudin C. 2015. Causal falling rule lists. Available online at http://arxiv.org/pdf/1510.05189.pdf (accessed September 8, 2017).
22. Chakarov A, Nori A, Rajamani S, et al. 2016. Debugging machine learning tasks. Available online at http://arxiv.org/pdf/1603.07292.pdf (accessed September 8, 2017).
23. Petrik M, Luss R. Interpretable policies for dynamic product recommendations. In: Proceedings of Conference on Uncertainty Artificial Intelligence, Jersey City, NJ, 2016, p. 74.
24. Provost F, Fawcett T. Robust classification for imprecise environments. Mach Learn. 2001;42:203–231.
25. Davenport MA, Baraniuk RG, Scott CD. Tuning support vector machines for minimax and Neyman-Pearson classification. IEEE Trans Pattern Anal Mach Intell. 2010;32:1888–1898.
26. Hajian S, Domingo-Ferrer J. A methodology for direct and indirect discrimination prevention in data mining. IEEE Trans Knowl Data Eng. 2013;25:1445–1459.
27. Feldman M, Friedler SA, Moeller J, et al. Certifying and removing disparate impact. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 2015, pp. 259–268.
28. Barocas S, Selbst AD. Big data's disparate impact. California Law Rev. 2016;104.
29. The U.S. EEOC. 1979. Uniform guidelines on employee selection procedures.
30. Varshney KR, Prenger RJ, Marlatt TL, et al. Practical ensemble classification error bounds for different operating points. IEEE Trans Knowl Data Eng. 2013;25:2590–2601.
31. Attenberg J, Ipeirotis P, Provost F. Beat the machine: challenging humans to find a predictive model's "unknown unknowns." ACM J Data Inf Qual. 2015;6:1.
32. Weiss GM. Mining with rarity: A unifying framework. SIGKDD Explorations Newsletter 2004;6:7–19.
33. Sahuguet A, Krauss J, Palacios L, Sangokoya D. Open civic data: Of the people, by the people, for the people. Bull Tech Comm Data Eng. 2014;37:15–26.
34. Shaw E. 2015. Improving service and communication with open data: A history and how-to. Ash Center, Harvard Kennedy School, Tech. Rep. Available online at http://datasmart.ash.harvard.edu/news/article/improving-service-and-communication-with-open-data-702 (accessed August 31, 2017).
35. Kapoor S, Mojsilović A, Strattner JN, Varshney KR.. From open data ecosystems to systems of innovation: A journey to realize the promise of open data. In: Proceedings of Data for Good Exchange Conference, New York, NY, 2015.
36. Schirner G, Erdogmus D, Chowdhury K, Padir T. The future of human-in-the-loop cyber-physical systems. Computer. 2013;36–45.
37. Kassahun Y, Yu B, Tibebu AT, et al. Surgical robotics beyond enhanced dexterity instrumentation: A survey of machine learning techniques and their role in intelligent and autonomous surgical actions. Int J Comput Assist Radiol Surg. 2016;11:553–568.
38. Lin HC, Shafran I, Murphy TE, et al. Automatic detection and segmentation of robot-assisted surgical motions. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 802–810.
39. Lin HC, Shafran I, Yuh D, Hager GD. Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. Comput Aided Surg. 2006;11:220–230.
40. Reiley CE, Plaku E, Hager GD. Motion generation of robotic surgical tasks: Learning from expert demonstrations. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, 2010, pp. 967–970.
41. Shademan A, Decker RS, Opfermann JD, et al. Supervised autonomous robotic soft tissue surgery. Sci Translational Med. 2016;8:37ra64–337ra64.
42. Alemzadeh H, Chen D, Lewis A, et al. Systems-theoretic safety assessment of robotic telesurgical systems. In: Proceedings of International Conference on Computer Safety, Reliability, and Security, 2015, pp. 213–227.
43. Azimian H, Naish MD, Kiaii B, Patel RV. A chance-constrained programming approach to preoperative planning of robotic cardiac surgery under task-level uncertainty. IEEE Trans Biomed Health Inf. 2015;19:612–1898.
44. Rayej S. 2014. How do self-driving cars work? Available online at http://robohub.org/how-do-self-driving-cars-work/ (accessed September 8, 2017).
45. Lowy J. 2016. Driver killed in self-driving car accident for first time. Available online at www.pbs.org/newshour/rundown/driver-killed-in-self-driving-car-accident-for-first-time (accessed September 8, 2017).
46. Duchi J, Glynn P, Johari R. 2016. Uncertainty on uncertainty, robustness, and simulation. SAIL-Toyota Center for AI Research, Stanford University, Tech. Rep.. Available online at http://aicenter.stanford.edu/uncertainty-on-uncertainty-robustness-and-simulation (accessed August 31, 2017).
47. Zhu Y, Janapa Reddi V. Cognitive computing safety: The new horizon for reliability. IEEE Micro. 2017;37:15–21.
48. Koopman P, Wagner M. Challenges in autonomous vehicle testing and validation. SAE Int J Transportation Saf. 2016;4:2016-01-0128.
49. Brynjolfsson E, Hitt L, Kim H. Strength in numbers: How does data-driven decision-making affect firm performance? In: Proceedings of International Conference on Information System, Shanghai, China, 2011, p. 13.
50. Singh M, Varshney KR, Wang J, et al. An analytics approach for proactively combating voluntary attrition of employees. In: Proceedings of IEEE International Conference on Data Mining Workshops, Brussels, Belgium, 2012, pp. 317–323.
51. Wei D, Varshney KR. Robust binary hypothesis testing under contaminated likelihoods. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, Australia, 2015, pp. 3407–3411.
52. Malioutov DM, Varshney KR. Exact rule learning via Boolean compressed sensing. In: Proceedings of International Conference Machine Learning, Atlanta, GA, 2013, pp. 765–773.
53. Gerard H, Rao K, Simithraaratchy M, et al. Predictive modeling of customer repayment for sustainable pay-as-you-go solar power in rural India. In: Proceedings of Data for Good Exchange Conf., New York, NY2015.
54. Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a'right to explanation'. In: Proceedings of ICML Workshop Human Interpretability, New York, NY. 2016, pp. 26–30.

Cite this article as: Varshney KR, Alemzadeh H (2017) On the safety of machine learning: cyber-physical systems, decision sciences, and data products. *Big Data* 5:3, 246–255, DOI: 10.1089/big.2016.0051.