AI & SOCIETY (2023) 38:685–696
https://doi.org/10.1007/s00146-022-01591-z

OPEN FORUM

Toward safe AI

Andres Morales-Forero[1] · Samuel Bassetto[1] · Eric Coatanea[2]

Received: 30 July 2021 / Accepted: 24 March 2022 / Published online: 23 November 2022

Abstract
Since some AI algorithms with high predictive power have impacted human integrity, safety has become a crucial challenge in adopting and deploying AI. Although it is impossible to prevent an algorithm from failing in complex tasks, it is crucial to ensure that it fails safely, especially if it is a critical system. Moreover, due to AI's unbridled development, it is imperative to minimize the methodological gaps in these systems' engineering. This paper uses the well-known Box-Jenkins method for statistical modeling as a framework to identify engineering pitfalls in the adjustment and validation of AI models. Step by step, we point out state-of-the-art strategies and good practices to tackle these engineering drawbacks. In the final step, we integrate an internal and external validation scheme that might support an iterative evaluation of the normative, perceived, substantive, social, and environmental safety of all AI systems.

Keywords Safe AI · Explainable AI · Interpretable AI · Trustworthy AI · Box-Jenkins framework · Ethical AI · Responsible AI · Relational validity

1 Introduction

Artificial Intelligence (AI) has performed exceptionally strongly in recent decades. Automatic and online image recognition, translators, driverless cars, and games have reached accuracy levels, in some instances, that surpass human levels (Mnih et al. 2015). However, the safety of some AI technologies has been questioned. For example, a self-driving vehicle killed a pedestrian in Arizona, AI-based prosecution systems have kept innocent people in jail, and IBM's automatic cancer treatment advisor made unsuitable recommendations (Marcus and Davis 2019, p. 11). Concerned about the unfortunate consequences of engineering pitfalls and the potential impacts of the unregulated deployment of these intelligent systems, governments, private companies, and researchers are looking for strategies to ensure that AI can be safely deployed in different economic sectors. The unbridled use of AI can be beneficial, but it also creates new risks for the environment, humans, and other sentient beings. Therefore, research on safe AI is urgently needed and a consensus must be reached on implementing new intelligent systems and validating those already in place, especially in the case of systems with a high potential impact on the community.

The latest AI developments are grounded in Machine Learning (ML), especially Deep Learning (DL). The complex structure of the most powerful ML models makes them impossible to interpret, turning them into black boxes. In addition to skepticism about decision-making, the opacity of these models has been known to conceal unfair decisions (Rudin 2019), thus generating mistrust about their safety. Moreover, naive errors in data processing and other kinds of negligence in adopting these models have demonstrated their limited substantive safety (Buolamwini 2017). In this article, we remind ML practitioners that ML models belong to a particular family of statistical models; therefore, frameworks such as the well-known Box-Jenkins framework used to adjust and validate statistical models can help ML modeling as well. This Box-Jenkins framework has been part of the "manual" of good practices in the statistical community for several decades and could well be applied in the AI community too. Using this framework, this paper presents a literature review identifying engineering pitfalls and factors that compromise AI safety and presenting some state-of-the-art strategies and good practices to tackle these drawbacks.

✉ Andres Morales-Forero
andres.morales-forero@polymtl.ca

Samuel Bassetto
samuel-jean.bassetto@polymtl.ca

Eric Coatanea
eric.coatanea@tuni.fi

[1] Polytechnique Montréal, Montréal, Canada

[2] Tampere University, Tampere, Finland

On the other hand, although this paper does not intend to validate the Box-Jenkins framework, four phases of clinical AI-system validation (statistical, relational, pragmatic, and ecological), proposed by Cabitza and Zeitoun (2019) are integrated into it to highlight the importance of iterative internal and external safety validations. The last section summarizes the main findings of the literature reviewed and shares the authors' insights on how to move toward safe AI.

2 The Box-Jenkins framework for safe AI

Due to the significant deployment of AI in society, it is becoming more and more urgent to ensure that these technologies work safely. Therefore, new frameworks and regulations are consequently being designed: The European Commission (2019) presented its Ethics Guidelines for Trustworthy AI; The Executive Office of the President of the United States (2019) published the National AI R&D strategic in the same year, followed by the guiding principles for AI by the Government of Canada (2021) and the divulgation of the guidelines on AI ethics by the Ministry of Science and Technology (MOST) of China (2021). These policies and frameworks contain general requirements and ethical principles about transparency, privacy, fairness, and accountability. However, they do not go into much detail and do not explain how those concepts are interrelated. Private companies like Rolls-Royce (2021), Facebook (2022) or IBM (2022) have also published guidelines for a Trustworthy, Responsible and Explainable AI. They are also high-level aspects around the AI implementation disregarding a particular operation of the systems, like by the governments.

In contrast, the Trustworthy Artificial Intelligence Implementation (TAII) Framework proposed by Baker-Brunnbauer (2021) does go into mid-level details identifying the systemic relationships of ethics within organizations for their products and services. However, although it might be helpful for system design and validation interconnecting the organizational business model, it does not show a complete process of adjusting a model. Many other frameworks envision translating high-level principles into mid- or low-level concepts as proposed by Shneiderman (2020), Hibbard (2012), Abràmoff et al. (2020), or Cabitza and Zeitoun (2019). However, they ignore the intrinsic statistical nature that ML models contain or do not contain all the steps of the modeling process. The Box-Jenkins method, on the other hand, is an iterative procedure that goes through all the stages of statistical modeling, especially time series and regressions. It is easy to understand, has been used in different contexts (Box et al. 2015, p.15), and applies in just three phases: (1) Identification, (2) Estimation and validation, and (3) Application. The first phase refers to data prepossessing and model identification. Once the model is identified, its parameters are estimated, and then the best model is selected using suitable criteria, and its assumptions are statistically validated. The selected model can be used if the validation phase is satisfactory; otherwise, it must return to the first phase. The validation step consists of the verification of purely statistical assumptions.

To describe the context of AI system safety and integrate both internal and external validation, we connect the Cabitza and Zeitoun (2019) validation framework into the Box-Jenkins framework (see figure 1). *Internal validation* consists of evaluating these algorithms' performance in controlled scenarios, while *external validation* is carried out in real-world situations. The validation process is composed of four steps: (1) Statistical validation, (2) Relational validation, (3) Pragmatic validation, and (4) Ecological validation. *Statistical validation* might be the most widely used by AI practitioners and refers to the analysis of metrics such as accuracy, sensitivity, specificity, robustness, and consistency. *Relational validation* involves direct system users, such as physicians, inspectors, or other end-users who interact directly with the system. In this step, the system's usability is evaluated, and it provides the first view of human-machine interaction in experimental but close-to-reality settings. *Pragmatic validation* assesses the system's functionality in real-world conditions, and *ecological validation* is a longitudinal pragmatic validation involving all the players (e.g., patients in the case of medical diagnosis).

2.1 Data preparation

Data are crucial for developing AI technologies. Generally, thousands or millions of records are required to train AI models. These records, usually called examples in ML

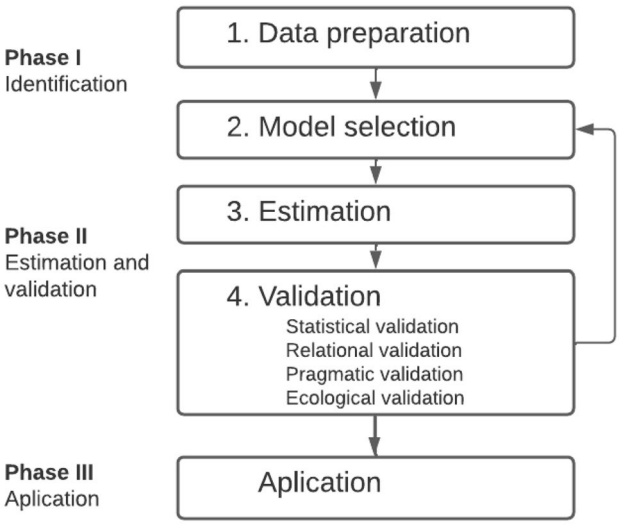

Fig. 1 The Box-Jenkins framework for safe AI

jargon, can come in different formats: cross-sectional or longitudinal data tables, images, videos, or voice recordings. Although it is impossible to look at all the records, it is essential to *become familiar with the data*. Not surprisingly, intelligent system designers must contextualize themselves before proposing solutions. Although this sounds obvious, ML practitioners often ignore the potential of descriptive statistics, multivariate analysis, or inspection of a few examples. For instance, some winning AI models unintentionally fooled the international PASCAL VOC competition because, to some extent, data were not preprocessed. The competition involves classifying realistic images with the highest possible accuracy level. Some of the accurate algorithms focused their attention on irrelevant photo areas to classify an image (Lapuschkin et al. 2016). The heat maps in figure 2 show two examples where a classifier might be focusing on irrelevant parts of the image. A ship was classified correctly due not to its morphology but to the presence of the sea. Horses were correctly classified since copyright text was on many images of horses. Perhaps this kind of error went unnoticed for years during the competition's history; it could have probably been foreseen if a preliminary analysis of the images had been carried out, especially in the case of the horses.

Alternatively, it is crucial to *pay attention to data collection details*. Wrong inferences might occur if sampling details are ignored. Even though collecting data that fully reflect reality might be impossible in some instances, it is essential to know the scope of these technologies' use. Representativeness, measurement errors, missing data, misclassification, high dimensionality, and unbalanced data affect the inferences and limit these models' use. Naturally, data is imperfect and unbalanced; some features are less prevalent than others. However, the precision of the algorithms must be guaranteed in some specific and essential cases. For example, driverless cars must detect pedestrians regardless of their skin color, or a medical diagnostic device for a specific disease must work for both in people with high and low socioeconomic status. In contrast, some IBM and Microsoft algorithms performed better on pale-male faces than on dark-skinned females (Buolamwini, 2017); ML-based clinical support might report inaccurate results in people with low socioeconomic status due to their underrepresentation in the health data (Gianfrancesco et al., 2018).

Excessive reliance on AI tools of this kind may amplify disparities and be unsafe. In cases where the performance of the algorithms cannot be guaranteed, these limitations must at least be reported to the users. A complete contextualization of the problem, the constant verification of the model's assumptions, sensitivity or robust analyses, and descriptive statistics help in the early identification of harmful features and uses of the model. As Buolamwini (2017) claims, this type of analysis should become "part of standard practice rather than merely a commendable option."

On the other hand, data integrity must be ensured, tested, and documented before a model is trained, according to The European Commission (2019). Moreover, when it comes to sensitive information, *privacy must be protected, and confidentiality preserved*.

2.2 Model selection

Data come in different structures such as longitudinal, unbalanced, or incomplete data. Modeling methodologies are proposed based on these characteristics. ML practitioners must remember that *the model must fit the data and not the data the model*. Ignoring the nature of data can lead to unreliable results, which work in theory but not in practice. For instance, the Google and Amazon search engines reflected gender discrepancies (Buolamwini 2017), even though methods already existed to remove gender stereotypes from embedding methods (Bolukbasi et al. 2016) such as the ones used by those search engines (Marcus and Davis 2019, p. 46). Class imbalance tends to generate biased learning and result in less predictive power in minority classes since most ML algorithms assume relatively balanced data (Zheng and Jin 2020). Weighting minority class distributions according to their learning difficulty level is a method that has been proposed to handle unbalanced data (He et al. 2008) and reweight for label-biased data (Jiang and Nachum 2020). Label-biased data might result in incorrect predictions, but optimization-based methods to control unfair predictions based on data that might favor certain classes have also been suggested in the literature (Thomas et al. 2019).

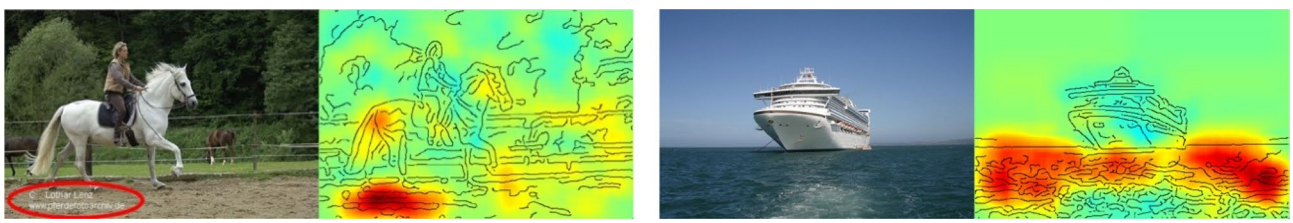

Fig. 2 Examples of anomalies detected in the classifiers of one of the PASCAL VOC competitions. © (2016) IEEE. Reprinted, with permission, from (Lapuschkin et al. 2016)

Sample size influences the scope of inferencing, data acquisition is often expensive, and ML algorithms might not generalize well with just a few training sessions. Learning from a few examples is essential in contexts such as anomaly detection because the scarcity of anomalies. The system would need to learn when a low-rate event is perceived. Methods such as few-shot learning (Fei-Fei et al. 2006; Fink 2005), prototypical networks (Snell et al. 2017), and matching networks (Vinyals et al. 2016) may be suitable to tackle these limitations since they require few data. The motivation underlying these algorithms is grounded in the idea that humans can learn new tasks quickly without a huge number of examples using prior knowledge. For example, a child who has seen a few photos of giraffes can identify a giraffe among multiple photos of animals. Transfer learning uses a convenient trained ML model to train a different model. With this method, the amount of data required is smaller than with a model trained from scratch. Other ways to gain in data include generating synthetic examples (Georgakis et al. 2017), data wrapping (Baird 1992), including expert knowledge (Vasconcelos and Vasconcelos 2017), and combining these techniques. Nevertheless, the safety of algorithms based on just a few examples should be tested since their inferences can be limited to particular cases, especially if they are trained on non-probabilistic samples.

AI algorithms have shown great predictive capacity, and thus statistical reliability. However, reliability is not the same as safety. If a critical system fails, it should fail safely, and other alternatives need to be available to the user. Thus, controllability is essential for a safe AI system. To control a system, users need to interact with the machine, and to interact with it, users need to understand its outputs: they should understand why a decision is made or how the machine performs its process. Therefore, *safe systems require explainable algorithms*. However, explainability might be the Achilles heel of the most capable AI algorithms. DL has emerged with a set of accurate but uninterpretable models that in some tasks have even surpassed the human level (Mnih et al. 2015). Because of their internal complexity, their underlying decision mechanism is incomprehensible, which makes users skeptical. This issue is known as the black-box phenomenon in AI. High-stakes decisions based on these black-boxes can harm society (Varshney and Alemzadeh 2017). For instance, a physician might not be able to correct an AI-based diagnosis (Akatsuka et al. 2019) or prevent a false positive. Accurate black-box systems may be grounded in spurious correlations (Lapuschkin et al. 2019), leading to unsafe decisions. Now more of those systems are being hastily produced, so the AI community needs to deliberate about AI safety and trustworthiness.

Making automatic decisions understandable or interpreting the mechanism inside these black boxes increases a system's safety perception, trust, and acceptance (Shin 2021), which are key factors to externally in ensuring that systems are externally valid. Explainable AI methods are developed with two schools of thought: (1) some researchers work on inherently interpretable models, while (2) others develop methods that are explainable post hoc. The first school creates ML models with transparent, easily understandable internal mechanisms, whereas the second creates methods to explain ML algorithms that are already trained. There are also hybrid approaches that combine both approaches.

Purely transparent models are perhaps the simplest AI algorithms; their structures allow their complete decision mechanism to be interpreted. Some examples are regression models, logistic models, and Bayesian models. The structure of these models is clear, so users can easily interpret and interact with model parameters to understand why and how the model decides. Although they have been widely used in contexts such as health (Mor-Yosef et al. 1990) and education (Kobrin et al. 2011), their statistical accuracy has been outperformed by more complex models. These include decision trees, which are robust, transparent models but they become less comprehensible to humans as the number of their nodes increases, except in some pathological models (Maimon and Rokach 2014, p. 53). Transparent models allow for complete interpretation, resulting in less uncertainty for users. However, methods such as Concept Whitening (CW) (Chen et al. 2020), deepLIFT (Shrikumar et al. 2019), and Grad-CAM (Selvaraju et al. 2019) sacrifice complete interpretability for more accurate DL. These techniques are model-specific approaches to interpret deep neural networks (DNNs). While CW, which is more associated with the first school of interpretable AI, introduces a mechanism to align concepts known to users with a latent space in the target DNN architecture during training, DeepLIFT and Grad-CAM produce feature-based explanations of trained DNNs. Feature-based explanations mean that some relevant features are highlighted in the model input to explain its output (see figure 3). DeepLIFT assigns importance scores to the input features, backpropagating the contributions of all neurons in the network to every feature in the input. Grad-CAM, which is concept-based like CW, uses gradients to weight feature maps on images and produce visual explanations of any convolutional neural network–based model. Case studies have shown that Grad-CAM improves human performance to classify images accurately, reveals the trustworthiness of a classifier, and helps identify biases in the input data (Selvaraju et al. 2019).

In the family of post hoc methods, agnostic approaches obtain feature-based explanations without altering the architecture of the model. Agnostic methods are attractive since they ignore the black box's underlying structure. They do not need to open the black box, but they do require query access, that is, they need the output produced for a given data point. LIME is an agnostic, gradient-based

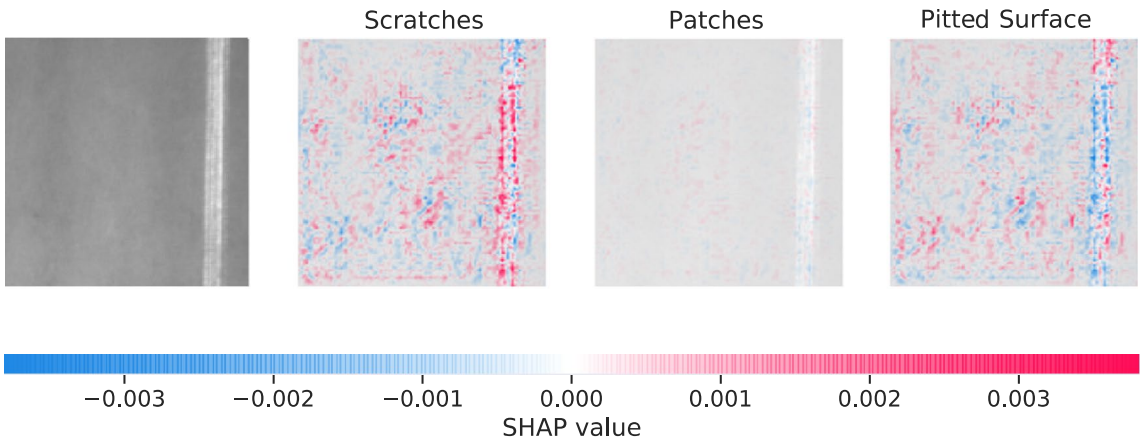

Fig. 3 Contrastive explanation of a type of defect based on SHAP values. The first image is taken from the NEU surface defect database (Song and Yan 2013; He et al. 2019; Dong et al. 2019). The three images at the right highlight the pixels that positively (red pixels) and negatively (blue points) affect the classifier's decision

strategy that explains any classifier or regressor's predictions faithfully and interpretably (e.g., a linear model) by approximating an interpretable model locally (Ribeiro et al. 2016). Like LIME, SHAP estimates feature attributions on individual instances, but it applies game theory to guarantee that the explanations satisfy specific properties (Lundberg and Lee 2017). SHAP outputs values that represent the features' contributions to predictions (see figure 3). Both SHAP and LIME provide visual explanations that can be easily interpreted by users.

Hybrid methods combine explainable models with black box models. Deep k-Nearest Neighbors (DkNN), Deep Weighted Averaging Classifier (DWAC), Self-Explaining Neural Network (SENN), and Contextual Explanation Networks (CENs) are some examples of hybrid approaches. DkNN combines k-nearest neighbor inferencing with the hidden representation of the training set in a DNN (Papernot and McDaniel 2018). Its nearest neighbors ensure each layer's interpretability since they represent a set of examples to explain a decision made by the DNN. DWAC makes predictions based on a weighted sum of training instances where the weights are determined by the distance from the instance to be evaluated to all training instances (Card et al. 2019). As in DkNN, the explanations for a decision are provided in terms of similar examples in the training set (the most highly weighted training examples). SENN uses locally simple interpretable models to generalize to more complex yet still interpretable models (Alvarez-Melis and Jaakkola 2018). SENN estimates these simple models' parameters via, for example, a DNN and regularizes their estimation to maintain interpretability. CENs learn to predict based on intermediate explanations. Those explanations are in the form of context-specific probabilistic graphical models. OPEN UP is a control chart used for online and post hoc fault diagnosis (Morales-Forero and Bassetto 2019).

There is still no consensus as to which of the two schools of thought provides valid explanations for the behavior of models. A known difference between inherently explainable models and post hoc methods is that interpretable models become more precise as they become more complex. Thus, the former is less accurate than the latter, which might be based on complex black boxes. This is not entirely true, though. If structured data with meaningful characteristics are available, models such as regression or decision lists can be as accurate as more complex models such as DNNs, boosted decision trees or random forests (Rudin 2019). The use of these approaches is context-dependent; verifying this trade-off and the data structure is a good practice for adopting these methods in particular cases. Nevertheless, assuming that interpretability might be sacrificed for performance, high-stakes decisions based on these black boxes can have harmful effects: using incomprehensible models, unknown shifts in the data would remain hidden, as well as wrong inferences based on incorrect patterns (Varshney and Alemzadeh, 2017). Furthermore, transparency and explainability are becoming mandatory in any deployment of AI-based critical systems (European Commission 2019; Executive Office of the President of the United States 2019; Ministry of Science and Technology (MOST) of China, 2021; Government of Canada 2021).

On the other hand, explanations depend on their users: while average users require practical, easy-to-digest, local explanations, technical experts may require more complex information, and managers more general explanations. Thus, *all the stakeholders' roles and knowledge need to be considered*. Users and designers must agree on what should be explained and how. They might be interested in feature-based explanations, concept-based explanations, contrastive explanations, or counterfactual explanations. As mentioned above, feature-based explanations highlight features in the

input to explain which ones are more relevant in predicting certain output; they can answer questions such as what areas in an input image influence the model decision. Their explanations are local: they apply to certain inputs. Concept-based explanations are grounded in known concepts, and thus are easier for users to understand. Contrastive and counterfactual explanations are quite similar: they both allow users to distinguish between different options in a system that answers questions such as "Why was a specific output not predicted?" or "What if something is changed in the model or the input?"

2.3 Estimation

In the ML community, the estimation process is called "training." Training is based on an objective function optimization using training data and is then tested with an independent data set that we call a test set. When testing with the test set, we expect that the model will produce similar results with unseen observations; in other words, that it will generalize well. Generalization is reached when the test set's error measure is as small as possible; that is, the model is accurate in the test set. The central assumption is that the training data and test data are drawn from the same distribution. However, in most practical cases, they come from related, not identical, distributions (Yao et al. 2020). Although it would be both necessary and sufficient for generalizing if an algorithm were to achieve "similar" performance on a test sample and a training sample that are "close," the condition of independent, identically distributed samples is still required (Xu and Mannor 2012). Therefore, special attention must be paid to the target population and sample mechanism for these training and test data. *Conclusions are limited to the target population with a proper sampling*.

In addition to explainability, perhaps the greatest weakness of AI models is the definition of their objective functions. The objective functions correspond to the mathematical representation of the instructions demanded by the system developers. The model parameters are estimated via the minimization or maximization of these functions. The idea is to formalize the designers' requirements based on these functions, but these requirements are often far from what is really wanted. For an AI to be safe, *it is important that what is asked of the machine corresponds precisely to what is wanted of the machine.* The objective functions might not fulfill all the needs, however. Complex tasks often require the learning process to be aligned with human values. A human-compatible system often requires more than the optimization of a single mathematical expression. Artificial General Intelligence (AGI) is an emergent topic that challenges researchers in this direction since it requires open-ended solutions: open-ended as a way to learn everything rather than to learn a specific task. The big challenge is formalizing open-endedness in mathematical expressions (Stanley 2019). Nevertheless, being aligned with human values does not guarantee the safety of the system, unforeseen events can occur even if a human operates the entire system. The problem is that machines do not have the common sense or goodwill to mitigate any unintended consequences. For instance, fully autonomous cars must immediately require solutions to ethical questions in situations such as described in the Trolley dilemma, where any result is undesirable and might pragmatically better express uncertainty. Some research has been done in this direction (Eckersley 2018; Beale et al. 2020).

The difficulty of expressing the correct objective functions is well illustrated in reinforcement learning. Two concrete problems related to objective functions' misspecification are negative side effects and reward hacking (Amodei et al. 2016). When the agent negatively affects the environment, this is a negative side effect; on the other hand, when the agent finds a simple or counterintuitive solution that maximizes the objective function but goes against the designer's desire, it is a case of reward hacking. In both cases, the agent learns to obtain the highest reward by applying strategies without common sense. These two problems were illustrated with the Coast Runners video game (Amodei and Clark 2016). Approaches to tackle this problem include penalizing changes in the environment and preventing the agent from reaching a risky position to deceive the system (Amodei et al. 2016). Letting the agent and the human work together to achieve the human's goals is another alternative. This approach is known as cooperative inverse reinforcement learning (Hadfield-Menell et al. 2016).

On the other hand, since *identifying vulnerabilities is crucial to ensure the system's safety*, techniques such as adversarial attacks and defense techniques have attracted increasing attention in the AI community. Adversarial attacks are algorithms specially designed to cause a malfunction in an ML model by perturbing its examples; they can affect the model both in the training stage and later on, usually during classification. The former is known as a poisoning attack and the latter as an evasion attack. Adversarial attacks can also be classified according to the level of knowledge: White-box attacks require full knowledge of the target model; black-box attacks only assume query access to the model, namely being able to obtain specific input predictions; finally, gray-box attacks have only limited access to the model. Since white-box and gray-box attacks have access to more information, they perform better than black-box attacks; thus, white-box adversarial attacks are the strongest (Ren et al. 2020). In this Box-Jenkins step, we will discuss poisoning attacks since they are the only ones that occur during the learning process. We will discuss evasion attacks in the next step, validation.

Poisoning attacks craft examples to affect the training process. When the adversary simply wants to reduce the

model's classification performance, these kinds of attacks are known as availability attacks (Biggio et al. 2012; Mei and Zhu 2015). In contrast, when the adversary seeks to misclassify a single target point, they are called targeted attacks (Koh and Liang 2017; Shafahi et al. 2018). Backdoor attacks, in contrast, are attacks in which the adversary controls one or more features to decrease the model's accuracy (Chen et al. 2017).

Defense techniques are alternative approaches to increase neural networks' robustness. These defense methods can either modify the data, modify the model, or use auxiliary information to increase the model's robustness. The methods that modify the data include adversarial training, gradient hiding, blocking transferability, data compression, and data randomization; methods that modify the model include regularization, defensive distillation, feature squeezing, deep contractive network, and mask defense. The Defense-GAN, MagNet (Meng and Chen 2017), and High-level Representation Guider Denoiser methods use auxiliary information. However, since white-box attacks have full access to the model, especially its parameters, they can adapt and craft adversarial samples directly for the target model. Many defense methods are vulnerable to such attacks, even when they are effective against black or gray-box attacks (Ren et al. 2020).

2.4 Validation

In this step, a validation process for implementation and safety monitoring of AI systems is described, mostly for critical systems. The procedure was initially proposed by Cabitza and Zeitoun (2019) for clinical contexts and consists of four stages: statistical validation, relational validation, pragmatic validation, and ecological validation. Some methodologies that aim to validate the model's safety at each stage are described below.

2.4.1 Statistical validation

Although a kind of validation is often done during the estimation process, it is necessary to evaluate performance with a completely independent data set to confirm that the algorithm does not overfit the training data. Roughly, the same objective function used for the training is also evaluated in the test set; if this value is close to the one obtained in the training process, the model is ready to use. In classification, the F score and the area under the receiver operating characteristic (ROC) curve are also commonly used to express statistical validity in terms of true-positive rates (sensitivity, recall, or probability of selection) and false-positive rates (1—specificity). These types of validation are the most common in the development of ML models. However, to ensure safety in this step, additional statistical validations might be required. For example, let us examine some evasion attack techniques. As mentioned above, evasion attacks are a kind of adversarial attack against a trained model. An evasion attack carefully creates malicious examples that are likely to be misclassified by the model. There are many kinds of evasion attacks, and new ones are constantly being invented. Detecting new adversarial observations can be more efficient than preventing them (Koo et al. 2019). Detecting corrupt observations can either robustify the model by retraining it based on the updated adversarial repository or prevent it from tampering with the system. Essentially, these adversarial detectors are either another binary ML classifier (Hendrycks and Gimpel 2016; Fidel et al. 2019; Koo et al. 2019) or a statistical test (Grosse et al. 2017) that classifies a test example as benign or adversarial. MagNet is another approach that, in addition to detecting, automatically corrects the corrupt sample so that it does not affect the system (Meng and Chen 2017). However, there are also special attacks that learn to fool both the classifier and the adversarial detector (Carlini and Wagner 2017). Hence, all types of adversarial defenses might be required. Some industries might require more sophisticated, up-to-date defense systems than others, such as banks or insurance companies versus manufacturing companies.

2.4.2 Relational validation

Achieving high accuracy is essential, but it might not be the most relevant result. Even if a system has been proven to be statistically valid, statistical validity is not enough to ensure its usability, especially in critical tasks (Cabitza and Zeitoun 2019). For instance; doctors can become less vigilant and over-rely on a system with high predictive capacity, even when it is wrong (Parikh et al. 2019); inspectors can provide correct diagnostics based on a system that provides right answers in 7 out of 10 cases if they are able to ignore the incorrect diagnostics interpreting the system outputs. Relational validation evaluates end-user–machine interactions. It validates whether the system is efficient and works reasonably well in real-world, but controlled settings (Cabitza and Zeitoun 2019). While statistical validation is focused on efficacy, relational validation is focused on the system's usability. ISO standards can help to specify and measure system usability (ISO 9241-110, 2020). Specifically, the recently published ISO/IEC TR 24028 (2020) provides an overview of topics relevant to building AI systems' trustworthiness. Trustworthiness, in the sense of the ability "to meet stakeholder expectations in a verifiable way," is the primary objective. Considerations of availability, resiliency, reliability, accuracy, safety, security, and data privacy are also included in this ISO.

On the other hand, users' perceived safety is a relevant factor in adopting technology. Perceived safety refers to "the

user's perception of the level of danger when interacting with a robot, and the user's comfort level during the interaction" (NíFhaoláin et al. 2020). Although human–machine interaction is a challenge for all automated systems, explainability allows human–machine communication in AI systems. Explainability enables the user to interact with the AI systems, providing information about a determinate decision; in certain instances, the user can arbitrarily modify the inputs and contrast the outputs, so the interaction is in both directions: machine-human and human–machine. Robot-human interaction, stress, affective state, and reliability have been measured to evaluate intelligent systems' safety. Physiological sensors, questionnaires, and direct input devices are often used after users have interacted and accumulated some experience with the system (Cabour et al. 2021). Although there has been immense progress in Explainable AI and assessment of its explanations, there is still much room for improvement.

2.4.3 Pragmatic validation

This kind of validation is similar to relational validation, but it is done later in the process. To pragmatically validate the system, it should be exposed to real-world conditions in an uncontrolled environment. Here, the objective is to evaluate the system's actual effectiveness.

Before exposing users to the system, *compliance with all rules and regulations must be ensured*. The concept of normative safety becomes important here; a system achieves normative safety if it meets all the relevant norms, regulations, and standards. In addition to all the legal or technical requirements that every system needs to meet according to its context and industry, special new regulations have emerged for some AI technologies (e.g., European Commission (2019)).

Factors such as transparency, explainability, fairness, governance and adaptability are common requirements in all these regulations, however it is not yet clear what level of explainability or transparency these systems require. The evaluation of the impact of these systems would be easier if counterfactual scenarios can be assessed but those type of analysis are limited given the complex architecture of the NN. Complete transparency allows assessing the causality of the inputs, i.e., it enables decision makers to fully understand the changes in the results due to small changes in the input data. Although some explainable algorithms such as SHAP or LIME have been proposed based on certain causality (Lundberg and Lee 2017), those methods remain exploratory, and sometimes inconsistent. The opaque nature of DNN, for instance, can still hide unknown shifts in the data and might provide wrong inferences based on incorrect patterns (Varshney and Alemzadeh 2017). Building inherently explainable models is more responsible since it is easier to evaluate the model causality (Rudin 2019). The interpretability of AI models is definitely what sets them apart from any other system, so by closing this gap the effectiveness of these AI-based systems can be evaluated regardless. Therefore, we believe that when it comes for safety in high-stake decisions, *interpretability must clearly be required for the current frameworks and regulations*.

2.4.4 Ecological validation

Ecological validation refers to longitudinal validation of the environmental benefits the system generates, that is, an assessment of the system's social and environmental fit over time (Cabitza and Zeitoun 2019). Factors such as throughput rates, workflow improvement, and net savings are validated here, along with cost-effectiveness analyses and indirect-user safety assessments. *Public, social, and environmental safety must be considered from the system design stage*, however. This stage is only a validation step to confirm what has been planned. Since the nature of the automated task might be dynamic, adjustments in the system may be needed. Therefore, continuous monitoring of these safety factors is recommended.

This kind of validation is important for ensuring that the AI system will not create unexpected social situations. Algorithmic impact assessments (AIAs) are frameworks that allow public and private agencies to evaluate the impact of automated decision systems in the environmental, privacy, and human rights policy domains. AIAs are becoming mandatory in any deployment of automated decision-making. Be as open as possible, allow auditing, ensure explainable algorithms, and preserve privacy are some of the common features of these AIAs. Alternative frameworks (e.g., the Aletheia Framework designed by Rolls-Royce (2021) or the Canadian guiding principles to ensure the effective and ethical use of AI published by The Government of Canada (2021)) provide essential measurements and general guidance on implementing AI systems.

"The core issue is that current AI systems mimic input data without regard either to social values or to the quality or nature of the data" (Marcus and Davis 2019, p. 47). For instance, Kim et al. (2019) reported that only 6% out of 516 published studies in medical imaging diagnosis externally validated their AI algorithms. Statistical accuracy indicators are certainly one of the main concerns of ML practitioners, suggesting indifference to questions such as: How compatible is the system with society, the environment, and sentient life? How well is it aligned with the social values? What are the risks to society and to nature of implementing the system? Although these questions might have subjective answers, methods to deal with biased or unfair decision-making are being developed. For instance, the Seldonian method based on constrained optimization allows one to

control undesirable behaviors in the predictions of ML models (Thomas et al. 2019). The constraints are measures to control unfair predictions probabilistically and intuitively; an undesirable behavior could be, for example, large differences in mean prediction errors for GPAs of applicants of different genders or for expected financial loss. Thomas et al. (2019) show how gender inequities in GPA predictions can be controlled and still produce more accurate estimates than state-of-the-art fairness-aware algorithms (e.g., Fairlean (Agarwal et al. 2018) and Fairness constraints (Zafar et al. 2017)). In contrast, Eckersley (2018) suggests that high-stakes decisions, such as medical decision support systems or autonomous weapons or risk assessment in criminal justice contexts, cannot be based only on strict mathematical objective functions. Otherwise, it would be impossible to make a good prediction "without violating strong human ethical intuitions"; ethical questioning, commonly there are no close-ended answers to many ethical questions, so these critical systems should exhibit some uncertainty in these cases. Human assistance is still required.

On the other hand, Deepfakes that involve the use of AI models to generate fake audio or visual content have shown their impact on public safety. Researchers have also increased their efforts to simulate these malicious activities and create defenses against them. Poisoning attack simulations to make discriminatory decisions (Solans et al. 2020); methods to detect fake visual content using physical or physiological aspects (Li et al. 2018; Yang et al. 2019; Agarwal et al. 2019), and capturing specific artifacts (Zhou et al. 2017, 2018) are products of those efforts.

The impact of these technologies on society and nature is undoubtedly an essential element of their implementation. Although legislation and frameworks already exist to prevent it from producing environmental, economic, and human damage (European Commission 2019; Executive Office of the President of the United States 2019; Ministry of Science and Technology (MOST) of China, 2021; Government of Canada 2021), there are still globally diversified technologies in which the safety factor takes a backseat, such as the very recent case of Facebook, which is accused of causing detrimental to the mental health of teenagers and being dishonest in its fight against hate and misinformation CBC (2021). Throughout this article, we have identified some factors and strategies for developing, validating, and monitoring AI. Still, there seems to be an endless discussion of safe AI while the misuse and abuse of these technologies continue.

3 Conclusion

In this article, we have used the Box-Jenkins framework to structure an analysis of factors and strategies to ensure safer adoption of AI technologies. Engineering pitfalls, plausible state-of-the-art solutions, and challenges have been identified throughout the model adjustment process, from data collection to external validation, showing the long road that lies ahead to ensure safe AI. Many of the new models are developed by big technology companies, which collect large volumes of information through social media or electronic devices. However, the ML community has given little thought to data collection, possibly since it has traditionally thought that the more, the better. Oversampling niche populations does not lead to a better understanding of the general population. Each element in the population of interest must have a greater than zero probability of being selected; otherwise, inferences concerning the entire population might be biased. Data collection spearheads the modeling process; therefore, probabilistic sampling methods need to be one of the first concerns for many ML practitioners. Sampling depends on the context and affects inferencing. Inferencing is relevant since real-life system performance can be affected, gender discrepancy can be produced, and the integrity of people's lives can be negatively impacted, to mention just a few issues.

We have also described how data preprocessing can bring to light collection gaps, contextualize the designers, and customize the modeling. Although powerful AI models exist in the literature, there are no one-size-fits-all models. Context-dependent modeling supports identification of a proper model and its validity. Internal and external validations must be performed to assess whether the system fulfills the task requirements and identify risks and possible vulnerabilities in all the safety dimensions: normative, perceived, substantive, social, and environmental safety.

The definition of objective functions represents a real challenge for AI practitioners. Parameter estimation based on simple target functions in close-ended environments might not be enough to ensure system safety. On one hand, obtaining objective parameter estimates can be very difficult since data scientists and software developers may embed their own task perceptions, and complex tasks might require more than a single mathematical expression. On the other hand, although techniques such as adversarial testing help to make the system more robust, those techniques still tend to be narrow tasks tested in controlled environments. That is why it is so important to externally validate algorithms.

The community agrees that explainable AI is imperative for AI safety, especially for high-stakes decisions. Nevertheless, there is still no consensus about what to explain and how. Although the latest agnostic explainable models seem to provide meaningful explanations and to be more in line with the unbridled nature of AI, researchers have recently advised that we stop explaining black-box models since this practice can harm society and inherently interpretable models can provide more faithful interpretations. Inherently interpretable models may be less accurate than black

boxes but providing adequate, timely information will allow users to act correctly if the system fails. Therefore, inherently interpretable approaches might be the best candidates for safer AI. Nevertheless, explanations depend on system users and their needs. Some system users may require global and counterfactual explanations, while others require local and feature-based explanations. Thus, system design that embraces the knowledge and needs of all stakeholders could allow more users to detect and assess the system's failures on a timely basis.

Machines' lack of goodwill or common sense can lead to unfortunate situations. So, human-in-the-loop or human-assist systems might be safer than fully autonomous systems. Although being aligned with human values is a subjective concept, general ethical considerations can be incorporated into the system and constantly evaluated.

Note that no verification of the framework used in this paper was, or was intended to be, carried out since it was only used as a creative way to describe all the issues related to safety in AI systems. However, this idea could serve as inspiration for future research and case studies in AI.

Acknowledgements This research is supported by the Consortium for Research and Innovation in Aerospace in Québec (CRIAQ), funded by Mitacs Accelerate program. The findings and conclusions in this report are those of the authors.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

References

Abràmoff MD, Tobey D, Char DS (2020) Lessons learned about autonomous ai: finding a safe, efficacious, and ethical path through the development process. Am J Ophthalmol 214:134–142

Agarwal A, Beygelzimer A, Dudík M, Langford J, Wallach H (2018) A reductions approach to fair classification. In International conference on machine learning, pp 60–69

Agarwal S, Farid H, Gu Y, He M, Nagano K, Li H (2019) Protecting world leaders against deep fakes. In Cvpr workshops, pp 38–45

Akatsuka J, Yamamoto Y, Sekine T, Numata Y, Morikawa H, Tsutsumi K (2019) Illuminating clues of cancer buried in prostate mr image: deep learning and expert approaches. Biomolecules 9(11):673

Alvarez-Melis D, Jaakkola TS (2018) Towards robust interpretability with self- explaining neural networks. http://arxiv.org/abs/1806.07538. Accessed 29 Jan 2021

Amodei D, Clark J (2016) Faulty reward functions in the wild. https://openai.com/blog/faulty-reward-functions. Accessed 1 Jul 2021

Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Man´e D (2016) Concrete problems in ai safety. Retrieved 14 Mar 2020, from http://arxiv.org/abs/1606.06565

Baird HS (1992) Document image defect models. Structured document image analysis. Springer, New York, pp 546–556

Baker-Brunnbauer J (2021) Taii framework for trustworthy ai systems. ROBONOMICS J Autom Econ 2:17

Beale N, Battey H, Davison AC, MacKay RS (2020) An unethical optimization principle. R Soc Open Sci 7(7):200462

Biggio B, Nelson B, Laskov P (2012) Poisoning attacks against support vector machines. Retrieved 20 Feb 2021. https://arxiv.org/abs/1206.6389

Bolukbasi T, Chang K.-W, Zou JY, Saligrama V, Kalai AT (2016) Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in neural information processing systems. MIT Press, pp 4349–4357

Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) Time series analysis: forecasting and control. Wiley, New York

Buolamwini JA (2017) Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Cabitza F, Zeitoun J-D (2019) The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. Ann Transl Med 7(8):161

Cabour G, Morales A, Ledoux E´, Bassetto S (2021) Towards an explanation space to align humans and explainable-ai teamwork. Retrieved 25 Jan 2021. https://arxiv.org/abs/2106.01503

Card D, Zhang M, Smith NA (2019) Deep weighted averaging classifiers. proceedings of the conference on fairness, accountability and transparency, pp 369–378. Retrieved 28 Jan 2021. http://arxiv.org/abs/1811.02579. https://doi.org/10.1145/3287560.3287595

Carlini N, Wagner D (2017) Adversarial examples are not easily detected: Bypassing ten detection methods. In Proceedings of the 10th acm workshop on artificial intelligence and security, pp 3–14

CBC (2021) Whistleblower testifies facebook chooses profit over safety, calls for 'congressional action'. CBC News. https://www.cbc.ca/news/world/facebook-whistleblower-testifies-profit-safety-1.6199886. Accessed 18 Feb 2022

Chen Z, Bei Y, Rudin C (2020) Concept whitening for interpretable image recognition. Nat Mach Intell 2(12):772–782

Chen X, Liu C, Li B, Lu K, Song D (2017) Targeted backdoor attacks on deep learning systems using data poisoning. Retrieved 25 Jan 2021. http://arxiv.org/abs/1712.05526

European Commission (2019) Ethics guidelines for trustworthy ai. Retrieved from https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

Dong H, Song K, He Y, Xu J, Yan Y, Meng Q (2019) Pga-net: Pyramid feature fusion and global context attention network for automated surface defect detection. IEEE Trans Industr Inf 16(12):7448–7458

Eckersley P (2018) Impossibility and uncertainty theorems in ai value alignment (or why your agi should not have a utility function). Retrieved 20 Mar 2020. https://arxiv.org/abs/1901.00064

Executive Office of the President of the United States (2019) The national artificial intelligence r&d strategic plan. Retrieved from https://trumpwhitehouse.archives.gov/wp-content/uploads/2019/06/National-AI-Research-and-Development-Strategic-Plan-2019-Update-June-2019.pdf

Facebook (2022). Facebook's five pillars of responsible ai. https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/. Accessed 18 Feb 2022

Fei-Fei L, Fergus R, Perona P (2006) One-shot learning of object categories. IEEE Trans Pattern Anal Mach Intell 28(4):594–611

Fidel G, Bitton R, Shabtai A (2019) When explainability meets adversarial learning: Detecting adversarial examples using SHAP Signatures. http://arxiv.org/abs/1909.03418. Accessed 17 Dec 2020

Fink M (2005) Object classification from a single example utilizing class relevance metrics. In Advances in neural information processing systems, pp 449–456

Georgakis G, Mousavian A, Berg AC, Kosecka J (2017) Synthesizing training data for object detection in indoor scenes. Retrieved 01 Dec 2020. https://arxiv.org/abs/1702.07836

Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G (2018) Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 178(11):1544–1547

Government of Canada (2021) Responsible use of artificial intelligence (ai). Retrieved 04 Feb 2021. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html#toc1

Grosse K, Manoharan P, Papernot N, Backes M, McDaniel P (2017) On the (statistical) detection of adversarial examples. Retrieved 21 Feb 2021. https://arxiv.org/abs/1702.06280

Hadfield-Menell D, Russell SJ, Abbeel P, Dragan A (2016) Cooperative inverse reinforcement learning. In Advances in neural information processing systems. MIT Press, pp 3909–3917

Hallows R, Glazier L, Katz M, Aznar M, Williams M (2021) Safe and ethical artificial intelligence in radiotherapy–lessons learned from the aviation industry. Clinical Oncology, 34(2), 99-101

He Y, Song K, Meng Q, Yan Y (2019) An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. IEEE Trans Instrum Meas 69(4):1493–1504

He H, Bai Y, Garcia EA, Li S (2008) Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp 1322–1328

Hendrycks D, Gimpel K (2016) Early methods for detecting adversarial images. Retrieved 01 Dec 2020. https://arxiv.org/abs/1608.00530

Hibbard B (2012) Decision support for safe ai design. In: International conference on artificial general intelligence, pp 117–125

IBM(2022). Explainable ai. https://www.ibm.com/watson/explainable-ai?utmcontent=SRCWW&p1=Search&p4=43700064515261160&p5=p&gclid=Cj0KCQiApL2QBhC8ARIsAGMm-KHAqR9Gb S91U33HXTEtZKshdCJbM4Qw7D7aVFO6fyOAEgMAkFrc8aAuNFEALwwcB&gclsrc=aw.ds. Accessed 18 Feb 2022

International Organization for Standardization (2020a). Ergonomics of human-system interaction—Part 110: Interaction principles. Retrieved 3 May 2021. https://www.iso.org/obp/ui/#iso:std:iso:9241:-110:ed-2:v1:en

International Organization for Standardization (2020b). Information technology—Artificial intelligence—Overview of trustworthiness in artificial intelligence. Retrieved 3 May 2021. https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:24028:ed-1:v1:en

Jiang H, Nachum O (2020) Identifying and correcting label bias in machine learning. In International Conference on Artificial Intelligence and Statistics 702–712

Kim DW, Jang HY, Kim KW, Shin Y, Park SH (2019) Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol 20(3):405–410

Kobrin JL, Sinharay S, Haberman SJ, Chajewski M (2011) An investigation of the fit of linear regression models to data from an sat® validity study. ETS Res Rep Ser 2011(1):i–21

Koh PW, Liang P (2017) Understanding black-box predictions via influence functions. In International Conference on Machine Learning, pp 1885–1894

Koo J, Roth M, Bagchi S (2019) HAWKEYE: Adversarial Example Detector for Deep Neural Networks. http://arxiv.org/abs/1909.09938. Accessed 12 Feb 2021

Lapuschkin S, Waldchen S, Binder A, Montavon G, Samek W, Muller K-R (2019) Unmasking clever hans predictors and assessing what machines really learn. Nat Commun 10(1):1–8

Lapuschkin S, Binder A, Montavon G, Muller KR, Samek W (2016) Analyzing classifiers: Fisher vectors and deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2912–2920

Li Y, Chang M.-C, Lyu S (2018) In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. Retrieved 01 Dec 2020. https://arxiv.org/abs/1806.02877

Lundberg SM, Lee S-I (2017) A unified approach to interpreting Model predictions. In: I. Guyon et al. (Eds) Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf. Accessed 8 Oct 2020

Maimon OZ, Rokach L (2014) Data mining with decision trees: theory and applications. World scientific. 81

Marcus G, Davis E (2019) Rebooting ai: Building artificial intelligence we can trust. Pantheon

Mei S, Zhu X (2015) Using machine teaching to identify optimal training-set attacks on machine learners. In Proceedings of the aaai conference on artificial intelligence vol 29

Meng D, Chen H (2017) Magnet: a two-pronged defense against adversarial examples. In Proceedings of the 2017 acm sigsac conference on computer and communications security, pp 135–147

Ministry of Science and Technology (MOST) of China (2021) New generation artificial intelligence ethics specifications. Retrieved 4 Feb 2021. http://www.most.gov.cn/kjbgz/202109/t20210926177063.html

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533

Morales-Forero A, Bassetto S (2019) Case study: a semi-supervised methodology for anomaly detection and diagnosis. In 2019 ieee international conference on industrial engineering and engineering management (ieem). IEEE, pp 1031–1037. https://doi.org/10.1109/IEEM44572.2019.8978509

Mor-Yosef S, Samueloff A, Modan B, Navot D, Schenker JG (1990) Ranking the risk factors for cesarean: logistic regression analysis of a nationwide study. Obstet Gynecol 75(6):944–947

NíFhaoláin L, Hines A, Nallur V (2020) Assessing the appetite for trustworthiness and the regulationof artificial intelligence in europe. In: Proceedings of the The 28th irish conference on artificial intelligence and cognitivescience, dublin, republic of ireland, 7-8 december 2020. CEUR Workshop Proceedings

Nauck D, Kruse R (1999) Obtaining interpretable fuzzy classification rules from medical data. Artif Intell Med 16(2):149–169

Papernot N, McDaniel P (2018) Deep k-nearest neighbors: towards confident, interpretable and robust deep learning. http://arxiv.org/abs/1803.04765. Accessed 28 Jan 2021

Parikh RB, Obermeyer Z, Navathe AS (2019) Regulation of predictive analytics in medicine. Science 363(6429):810–812

Ren K, Zheng T, Qin Z, Liu X (2020) Adversarial attacks and defenses in deep learning. Engineering 6(3):346–360

Ribeiro MT, Singh S, Guestrin C (2016) "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. Association for computing machinery, New York, NY, USA, pp 1135–1144. Retrievedfrom https://doi.org/10.1145/2939672.2939778

Rolls-Royce. (2021). The aletheia framework. https://www.rolls-royce.com/sustainability/ethics-and-compliance/the-aletheia-framework.aspx. Accessed 1 July 2021

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2019) Grad- CAM: visual explanations from deep networks via gradient-based localization. http://arxiv.org/abs/1610.02391. https://doi.org/10.1007/s11263–019–01228–7. Accessed 25 Jan 2021

Shafahi A, Huang WR, Najibi M, Suciu O, Studer C, Dumitras T, Goldstein T (2018) Poison frogs! targeted clean-label poisoning attacks on neural networks. Retrieved 01 Dec 2020. https://arxiv.org/abs/1804.00792

Shin D (2021) The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. Int J Hum Comput Stud 146:102551

Shneiderman B (2020) Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. ACM Trans Interact Intell Syst (TiiS) 10(4):1–31

Shrikumar A, Greenside P, Kundaje A (2019) Learning important features through propagating activation differences. http://arxiv.org/abs/1704.02685. Accessed 21 Jan 2021

Snell J, Swersky K, Zemel R (2017) Prototypical networks for few-shot learning. In Advances in neural information processing systems. MIT Press, pp 4077–4087

Solans D, Biggio B, Castillo C (2020) Poisoning attacks on algorithmic fairness. Retrieved 20 Dec 2021. https://arxiv.org/abs/2004.07401

Song K, Yan Y (2013) A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. Appl Surf Sci 285:858–864

Stanley KO (2019) Why open-endedness matters. Artif Life 25(3):232–235

Thomas PS, da Silva BC, Barto AG, Giguere S, Brun Y, Brunskill E (2019) Preventing undesirable behavior of intelligent machines. Science 366(6468):999–1004

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

Vasconcelos CN, Vasconcelos BN (2017) Increasing deep learning melanoma classification by classical and expert knowledge-based image transforms. CoRR. http://arxiv.org/abs/1702.07025, 1

Vinyals O, Blundell C, Lillicrap T, Wierstra D, et al. (2016) Matching networks for one shot learning. In Advances in neural information processing systems. MIT Press, pp 3630–3638

Xu H, Mannor S (2012) Robustness and generalization. Mach Learn 86(3):391–423

Yang X, Li Y, Lyu S (2019) Exposing deep fakes using inconsistent head poses. In Icassp 2019–2019 ieee international conference on acoustics, speech and signal processing (icassp), pp 8261–8265

Yao L, Chu Z, Li S, Li Y, Gao J, Zhang A (2020) A survey on causal inference. Retrieved 18 Feb 2021. https://arxiv.org/abs/2002.02770

Zafar MB, Valera I, Rogriguez MG, Gummadi KP (2017) Fairness constraints: mechanisms for fair classification. In Artificial intelligence and statistics. PMLR, pp 962–970

Zheng W, Jin M (2020) The effects of class imbalance and training data size on classifier learning: an empirical study. SN Comput Sci 1(2):1–13

Zhou P, Han X, Morariu VI, Davis LS (2017) Two-stream neural networks for tampered face detection. In 2017 IEEE conference on computer vision and pattern recognition workshops (cvprw), pp 1831–1839

Zhou P, Han X, Morariu VI, Davis LS (2018) Learning rich features for image manipulation detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1053–1061