The relationship between trust in AI and trustworthy machine learning technologies

Ehsan Toreini
ehsan.toreini@ncl.ac.uk
School of Computing
Newcastle University
United Kingdom

Mhairi Aitken
mhairi.aitken@ncl.ac.uk
Business School
Newcastle University
United Kingdom

Kovila Coopamootoo
kovila.coopamootoo@ncl.ac.uk
School of Computing
Newcastle University
United Kingdom

Karen Elliott
karen.elliott@ncl.ac.uk
Business School
Newcastle University
United Kingdom

Carlos Gonzalez Zelaya
c.v.gonzalez-zelaya2@ncl.ac.uk
School of Computing
Newcastle University
United Kingdom

Aad van Moorsel
aad.vanmoorsel@ncl.ac.uk
School of Computing
Newcastle University
United Kingdom

ABSTRACT

To design and develop AI-based systems that users and the larger public can justifiably trust, one needs to understand how machine learning technologies impact trust. To guide the design and implementation of trusted AI-based systems, this paper provides a systematic approach to relate considerations about trust from the social sciences to trustworthiness technologies proposed for AI-based services and products. We start from the ABI+ (Ability, Benevolence, Integrity, Predictability) framework augmented with a recently proposed mapping of ABI+ on qualities of technologies that support trust. We consider four categories of trustworthiness technologies for machine learning, namely these for Fairness, Explainability, Auditability and Safety (FEAS) and discuss if and how these support the required qualities. Moreover, trust can be impacted throughout the life cycle of AI-based systems, and we therefore introduce the concept of Chain of Trust to discuss trustworthiness technologies in all stages of the life cycle. In so doing we establish the ways in which machine learning technologies support trusted AI-based systems. Finally, FEAS has obvious relations with known frameworks and therefore we relate FEAS to a variety of international 'principled AI' policy and technology frameworks that have emerged in recent years.

CCS CONCEPTS

• **Applied computing** → **Sociology**; • **Social and professional topics** → **Computing / technology policy**; • **Security and privacy** → *Human and societal aspects of security and privacy*; • **Computing methodologies** → **Artificial intelligence**; **Machine learning**.

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02…$15.00
https://doi.org/10.1145/3351095.3372834

KEYWORDS

trust, trustworthiness, machine learning, artificial intelligence

ACM Reference Format:
Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3351095.3372834

1 INTRODUCTION

Growing interest in ethical dimensions of AI and machine learning has led to the focus on ways of ensuring trustworthiness of current and future practices (e.g. European Commission 2019, IBM n.d.). The current emphasis on this area reflects recognition that maintaining trust in AI may be critical for ensuring acceptance and successful adoption of AI-driven services and products [67, 81]. This has implications for the many AI-based services and products that are increasingly entering the market. How trust is established, maintained or eroded depends on a number of factors including an individual's or group's interaction with others, data, environments, services, products and factors, which combine to shape an individual's perception of trustworthiness or otherwise. Perceptions of trustworthiness impact on AI and consequently, influence a person's decision and behaviour associated with the service or product. In this paper, we research the connection between trust and machine learning technologies in a systematic manner. The aim is to identify how technologies impact and relate to trust, and, specifically, identify trust-enabling machine learning technologies. AI and machine learning approaches are *trustworthy* if they have properties that one is *justified* to place trust in them (see [7] for this manner of phrasing).

It is important to highlight the difference between studying trust in AI and studying ethics of AI (and data science). Trustworthy AI is related to normative statements on the qualities of the technology and typically necessitates ethical approaches, while trust is a response to the technologies developed or the processes through which they were developed (and may not necessarily - or entirely - depend on ethical considerations). Ethical considerations behind

the design or deployment of an AI-based product or service can impact perceptions of trust, for instance if trust depends on having confidence in the service not discriminating against the trusting entity (or in general). However, there may be cases where ethics is not a consideration for the trusting entity when placing trust in a service, or, more frequently, if ethics is one of the many concerns the trusting entity has in mind. In what follows, we will also see that trust-enhancing machine learning technologies can be related to various 'Principled AI' frameworks, such as Asilomar AI Principles introduced in 2017, Montréal Declaration for Responsible Development of Artificial Intelligence in 2018 and IEEE Ethically Aligned Design Document.

The aim of this paper is to identify the ways in which (classes of) machine learning technologies might enhance or impact trust, based on trust frameworks drawn from ethics, social sciences and computing and algorithm design literature on technological trust qualities. Figure 1 outlines our approach. At the centre of Figure 1 is the end product of this paper, trust-enhancing technologies in machine learning and their classification in technologies for Fairness, Explainability, Auditability and Safety (FEAS). The downwards arrows indicate that these technologies are derived from trust frameworks from social science literature (particularly organisational science). The upward arrow indicates that the FEAS-classification of technologies was informed by the various Principled AI frameworks that shape the ethics and policy discussion in many nations (this is discussed in Section 4.2).

As indicated in Figure 1, we base our discussion on the widely accepted ABI (Ability, Benevolence, Integrity) principles underlying trust, as introduced by Mayer et al.[67] and extended to include Predictability by Dietz and Den Hartog [29] (a.k.a ABI+). We add to this a temporal dimension, from initial trust to continuous trust, as discussed by Siau et al.[59]. This gives us a base to understand trust in general, and we augment this further by integrating Siau's perspective on trust in technology, which identifies that trust is impacted by Human, Environmental and Technological qualities (referred to as the technologies' HET qualities in what follows). We will discuss these steps to go from the ABI+ model to HET qualities of trustworthy technologies in Section 2.

To summarise, the contributions of this paper are as follows:

- We draw on social science literature, particularly from organisational science, to apply established principles of trust to examine the qualities for technologies to support trust in AI-based systems (primarily based on the ABI and ABI+ framework and the HET qualities).
- We identify how trust can be enhanced in the various stages of an AI-based system's life-cycle, specifically the design, development and deployment stages. We therefore introduce the concept of an AI *Chain of Trust* to discuss the various stages and their interrelations.
- We introduce a FEAS (Fairness, Explainability, Auditability, Safety) classification of machine learning technologies that support and enable trust and establish the relation between these trust-enhancing technologies and the HET qualities.
- We discuss how our technology classification and trustworthy machine learning techniques relate to various Principled AI framework considered by policy makers and researchers in ethics and associated topics.

2 TRUST

This section discusses trust frameworks we can use to classify and identify trustworthy machine learning technologies. It discusses the top half of the paper contribution provided in Figure 1, the box with Trust and with Trust in Technology. In Section 2.1 we introduce the ABI+ framework to describe trust in general, i.e., not restricted to trust in technology. Section 2.2 reflects on trust in technology and science, recognising that AI-based services are based on science and manifest as technologies. The final sections provide the framework developed by Siau, which includes a discussion on time-sensitivity of trust (Section 2.4) and recognises three types of qualities technologies may exhibit that impact trust (Section 2.3).

Trust is discussed across many diverse social science literature leading to an abundance of definitions and frameworks available through which to examine the concept. It is a concept which in everyday conversation is routinely and intuitively used and yet remains challenging to define and study. Andras et al. [5] summarise some of the ways that trust has been approached across different disciplines: "In the social world trust is about the expectation of cooperative, supportive, and non-hostile behaviour. In psychological terms, trust is the result of cognitive learning from experiences of trusting behaviour with others. Philosophically, trust is the taking of risk on the basis of a moral relationship between individuals. In the context of economics and international relations, trust is based on calculated incentives for alternative behaviours, conceptualised through game theory.' A comprehensive review of literature relating to trust is beyond the scope of this paper. Here we focus on established models to examine the nature of trust and discuss how this relates to technology.

2.1 The ABI+ Framework: Ability, Benevolence, Integrity and Predictability

The ABI framework introduced by Mayer et al [67] suggested that the three main attributes which will shape an assessment of the trustworthiness of a party are: Ability, Benevolence and Integrity. The model discusses the interactional relationship between a trustor (the entity that trusts) and a trustee (the entity to be trusted). Building on the work of Mayer et al [67], Dietz and Den Hartog [29] outline three forms of trust: trust as a belief; a decision and; an action. While many studies have focused on trust as a belief in isolation from actions, Dietz and Den Hartog [29] regard the three forms as being the constituent parts of trust which are most usefully examined together.

In the ABI model, ability is defined as the perception of: "that group of skills, competencies, and characteristics that enable a party to have influence within some specific domain". The *specific domain* is crucial as assessments of a party's ability will vary according to particular tasks or contexts. Benevolence is defined as "the extent to which a trustee is believed to want to do good to the trustor". To be considered to possess Integrity a trustee must be perceived to adhere "to a set of principles that the trustor finds acceptable". This requires confidence that the trustee will act in accordance to a set of principles and that those align with the values of the trustor.

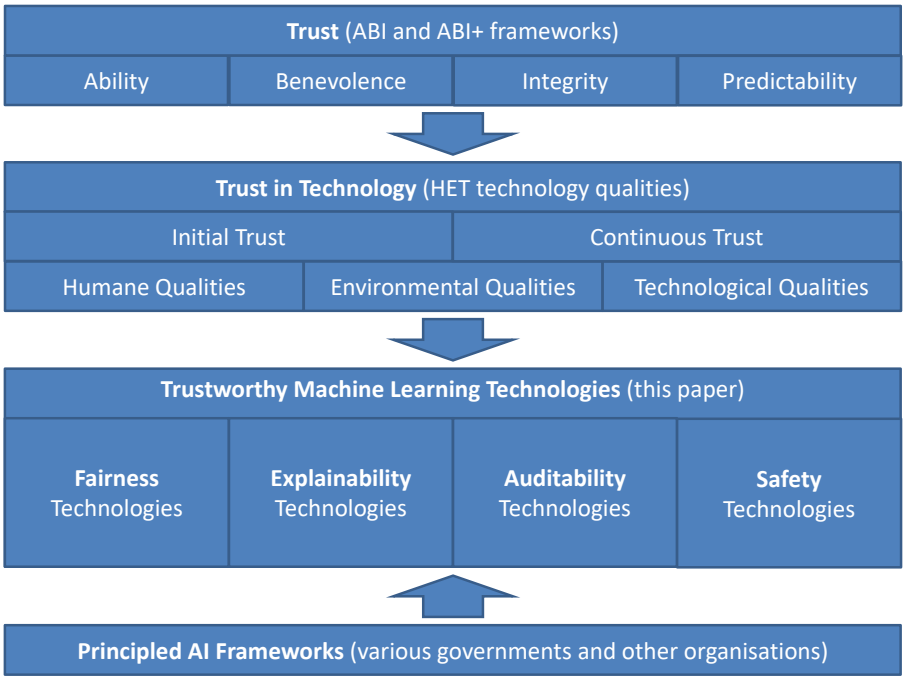

Figure 1: The contribution of this paper: identification of trust-enhancing Machine Learning technologies based on social sciences literature and relating these with Principled AI frameworks

Since its inception Mayer et al. [67]'s framework has been adapted and expanded to acknowledge the importance of Predictability or Reliability in shaping perceived trustworthiness. Dietz and Den Hartog [29] developed the ABI+ model suggesting that the four key characteristics on which judgements of trustworthiness are based are: Ability; Benevolence; Integrity and; Predictability. Predictability will reinforce perceptions of the Ability; Benevolence and; Integrity of the trustee. Considering the role of Predictability, draws attention to the importance of trust being sustained overtime through ongoing relationships.

While each of the attributes are related and may reinforce one another they are also separable [67]. One party may trust another even if they perceive one or more of these attributes to be lacking. As such trust -and trustworthiness- should not be thought of in binary terms but rather trust exists along a continuum.

As such evaluating one party's trust in another is more complex than a straightforward assessment of whether or not they trust that party and statements such as "A trusts B" are overly-simplistic, instead trust is described in statements reflecting the conditional nature of trust, for example: "A trusts B to do X (or not to do Y), when Z pertains…" [29, p. 564].

As well as being shaped by judgements of trustworthiness the act of trusting also depends on a range of external and contextual factors, personal attributes and traits of the *trustor*. An individual's predisposition or ideological position will impact on the extent to which they trust particular individuals/organisations, and these positions will shape how they receive, interpret and respond to information about the other party [29].

As the context changes, for example relating to cultural, economic, political or personal developments, so levels of trust and perceptions of trustworthiness also change. Therefore trust is characterised as an ongoing relationship rather than a static concept. Moreover, trust can be strengthened – or conversely weakened – through interactions between trustors and trustees: "outcomes of trusting behaviours will lead to updating of prior perceptions of the ability, benevolence, and integrity of the trustee" [67, p. 728]. See also Section 2.4 for a discussion of the time- -sensitive nature of trust.

2.2 Trust in Science and Technology

There is a significant body of literature in the field of Science and Technology Studies (STS) examining public relationships with science and technology, and, in particular, the role of trust in these relationships. In 2000, the UK House of Lords Science and Technology Committee published a landmark statement (which continues to be widely cited) stating that there was a 'crisis of trust in science'. This reflected wider discourses suggesting that a series of high-profile scientific controversies and scandals (e.g. BSE, thalidomide and the MMR triple vaccine), together with the rapid pace of scientific progress had resulted in an erosion of public trust in science [3].

This led to considerable attention directed at 'improving' public trust in science, typically through efforts to increase public understanding of science, on the assumption that, where the public is sceptical or mistrusting, this can be explained by ignorance or lack of understanding, and as such can be 'corrected' through better dissemination of scientific knowledge or *facts* [3]). However, such

approaches are now widely discredited as it is recognised that they overlook the role of members of the public in actively engaging with scientific knowledge rather than being "passive recipients of scientific knowledge" [24, p. 206]. Members of the public critically assess, deconstruct, question and evaluate claims to scientific knowledge in line with their own ideologies, experiences and the contexts in which the information is received [45]. This active process shapes people's trust as beliefs as well as informing the trust decisions and actions that are taken.

The public's relationship with science and technology is too sophisticated to be characterised by a simple trust/distrust binary relationship. Rather, in many cases the public adopts an ambivalent form of trust – described by Wynne [91–93] as an: *as if* trust. This takes account of the public's "knowingly inevitable and relentlessly growing dependency upon expert institutions" [93, p. 212].

With regard to AI-based technologies, the dependence on knowledge and behaviours of experts is clear, and trust is increasingly conditional. This implies that people do not automatically have confidence in particular innovations, scientists or scientific institutions (but equally lack of absolute trust does not mean that innovations will be met with public opposition). There can be dissonance between the trust beliefs held and the decisions and actions taken based on contextual, personal or organisational factors. In particular, even where people do not fully trust the technology they may use a service driven by AI if they feel there is no alternative option.

2.3 Trust Technology Qualities: Humane, Environmental and Technological

To further understand how technology interfaces with trust, Siau et al. [81] identify qualities of technologies that relate to trust and the concepts in the ABI+ framework of Section 2.1. The authors recognise three types of conditions to demonstrate the potential for a technology to be perceived as trustworthy: humane, environmental and technological qualities.

Humane Qualities. Humane qualities refer to the actions that attract individuals possessing a risk–taking attitude. The effectiveness of this quality depends on the personality type, past experiences and cultural backgrounds of the individuals. This is linked to the ability of the trustee to satisfy the curiosity of the trustor in testing a desired task. In other words, if cultural background resonates and if testing a product or service is feasible, this will typically enhance trust.

Environmental Qualities. Environmental qualities consider elements that are enforced by the qualities of the technology provider. First, it heavily relies on the nature of the task that the technology handles. The sophistication of the task has a potential to attract trustworthiness or cause distraction. The pattern of the establishment of trustworthiness differs in various places depending on the education system, level of their accessibility to novel modern advancements and subtle inherent cultural backgrounds. Yuki et al. [95] discussed such cultural impact on the trust establishment pattern. Institutional factors are another environmental parameter for trust. Siau et al. [81] defined it as "the impersonal structures that enable one to act in anticipation of a successful future endeavor". They collected two aspects for this concept: the institutional normality and structural assurance. The first deals with efficiency of the organisational structure and the later refers to the ability of an institution to fulfil the promises and its commitments.

Technological Qualities. Finally, technological qualities determine the capacity of the technology itself to deliver the outcome as promised. This commitment is multi-dimensional. First, the technology needs to yield the results efficiently. Thus, it needs to establish an agreed performance metric and assure its outcome yields into the desirable range of the metric. Second, the technology should define concrete boundaries for their solution. The user (as potential trustors) of the technology should be provided with enough information to infer the purpose of the technology and set their expectations sensibly based on such understanding. Lastly, The process of the technology outcome is another factor in trustworthiness. The technology should be able to reply to potential queries about *how* they concluded such outcome and *why* it led to the such performance. This aspect outlines the relation between the performance and purpose aspects.

2.4 Time-domain: Initial and Continuous Trust

A final element to support our understanding of trustworthy technologies is understanding trust as it develops over time, as also discussed in [59]. Highlighting the importance of predictability in the ABI+ model, the dynamics of relationship between trustor and trustee is an ongoing process. Usually, it requires initial preconditions to be satisfied, provided through first impressions, which is referred to as *initial trust*. After the initial phase, trust levels may change, for a variety of reasons, and this is referred to as *continuous trust* in the literature [81]. Typical examples that may impact continuous trust in AI-based services and products are data breaches, privacy leaks or news items on (unethical) business or other practices.

In what follows we will consider an additional time-sensitive element for trustworthy AI-based technologies, namely that of the service and product life cycle, both in terms of moving between the stages of design, development and deployment, as well as in terms of the machine learning pipeline, which includes data input, algorithm selection and output presentation. See the next section, and particularly Section 3.3.

3 MACHINE LEARNING AND THE CHAIN OF TRUST

In this section we introduce the concept of a Chain of Trust, which connects trust considerations in the stages of the machine learning pipeline and, when considered over time, the AI-based service or product may iterate through these stages (effectively expending the chain into a cycle, as illustrated in Figure 2). Before introducing the Chain of Trust in Section 3.3, we briefly review the basics of machine learning, as well as the notion of a machine learning pipeline.

3.1 Basics of Machine Learning

A machine learning algorithm is basically a function as $y = f_\theta(x)$ (with exception of a few algorithms such as nearest–neighbour [4]). In this equation, f_θ represents the function that maps input to the

output, i.e. the machine learning model. In this function, $\theta \in \Theta$ denotes a set of values that is tuned for the optimal operation of f. θ is calculated based on a pre-defined loss function that measures the similarity or dissimilarity between samples. The input of a model, x, correspond to a set of *features* which is a vector of values that represents to data, a.k.a. dataset. Finally, y represents the output of the algorithm to fulfil the task that it meant to undergo, i.e. *supervised*, *unsupervised* and *reinforcement* learning.

In supervised learning, the output y is meant to be an assignment of the input x to a pre-defined label set. Supervised Learning algorithms are mainly used in object recognition [54], machine translation [83], filtering spams [32], etc. For example, a fraud detection classifier would assign two labels (fraud or benign) to an input feature which is derived from a transaction.

Unsupervised Learning methods are used when the input x and output y are both unlabelled. In these methods, the task is to determine a function f_θ that takes x as input and detects a hidden pattern representation as output, y. The problems that unsupervised learning tackles include grouped a dataset based on a similarity metric (a.k.a clustering [48]), projecting data to a reduced dimension space (e.g. PCA methods [53]) and pre–training algorithms for the other tasks (i.e. pre–processing methods) [36].

Reinforcement Learning [84] methods maps x to a set of policies as y. In these techniques, f_θ determines an action, an observation or a reward (y) that should be taken into account when situation x is observed. Reinforcement learning techniques are concerned with how an *agent*, suppose a rescue robot, should behave under certain circumstances to maximise the chances of fulfilling a purpose.

3.2 Machine Learning Pipeline

Regardless of the task, the use of any machine learning algorithm implies activities in various stages, called the *pipeline*. This is depicted in Figure 2 through the chain of circles. First, one collects the data from a source and store the digital representation in a database ('Data Collection' in Figure 2) . Then, such data undergoes pre-processing methods to extract certain features, as x in the machine learning equation, labelled 'Data Preparation' and 'Feature Extraction', respectively.

When x is ready to be processed for the supposed task, the features are divided into at least two groups to attain two purposes. The first group of features ('Training' in Figure 2) are used to tune θ to optimise the output (y) of the function f. The group of features for this purpose is called the training data. The training process can be offline or online, depending on how static the training data is with respect to the life-cycle of a model. The online fashion deals with dynamic training in which the model re-tunes θ when new training data arrives. In contrast, in offline mode, the training stage only operates once on a static training dataset.

The second group of features is used for the purpose of verifying the generalisation of the f_θ parameters when the model faces an unknown parameters when new training data appears in the process ('Testing' in Figure 2). Verification is done by assessing the efficiency of the performance through some chosen metric. For example, in classification, a typical accuracy metric is the proportion of the test data that has been mapped to their original label correctly by f_θ. The features and data used at the testing stage are called *test data*. When the model passes the verification stage with a sufficiently good performance, then they are applied in the wild. This stage is known as 'Inference', in which the trained model is deployed to face unseen data. In this context, θ is fixed and y is computed for $f_\theta(x)$ when x is unknown (i.e. not contained in either test and training datasets).

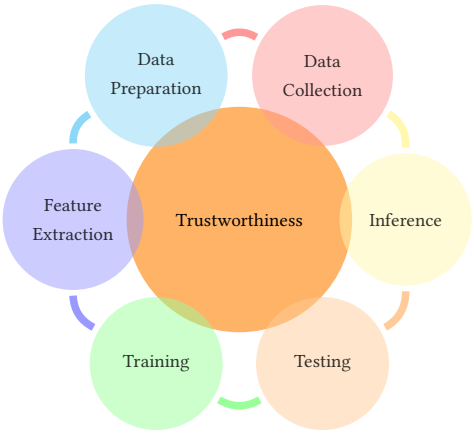

Figure 2: Various stages in the machine learning pipeline, motivating the notion of a Chain of Trust.

In a real–world context, a machine learning system is designed and implemented for a certain use case. Let us consider face recognition as an example of how the pipeline functions. The recognition system is based on a machine learning algorithm, $f_\theta(x)$, that distinguishes facial properties and maps them on a pre–defined set of authorised people. This is the supervised learning task. Before the deployment of the system, a large number of facial samples is collected, selected and prepared for the system. For instance, images with corrupt or blurred faces may be removed or the lightening of the images is re-balanced. This is done in the data collection and data preparation stages of the machine learning pipeline. The input to the algorithm, x, is a set of features derived from the facial image samples using image processing techniques. This is the feature extraction stage. Then, the algorithm is trained (optimising θ) with a set of pre–labelled facial images from the sample collection, the training set. After that, in the testing stage, the trained algorithm is tested with another unknown set of facial image samples verifying the accuracy of its recognition. Finally, in the inference stage, the trained and tested system is deployed in a real-life setting, categorising images of 'new' faces captured in real–time.

In what follows we group some of the stages in the machine learning pipeline. The first group of stages concerns data-centric aspects, involving data collection methods, pre-processing techniques and extraction of useful features for the analysis. The second group of stages is model-centric, the stages in Figure 2 that deal with the tuning the model to the best performance ('Training' stage), evaluating the trained model for confirmation of the desirable performance ('Testing' stage) and deployment of the model for the real-world application ('Inference' stage).

3.3 Chain of Trust

We are now in a position to introduce the notion of Chain of Trust. Based on the machine learning pipeline depicted in Figure 2 it becomes clear that technologies may impact trust in the resulting service or product in various stages of the pipeline. For instance, better methods to clean the data during 'Data Preparation' may avoid bias in the output of algorithms, which in turn helps to enhance trust once it becomes visible to users or the public. There are a number of important dimensions to the Chain of Trust, each of which demonstrates the importance of continuous trust as discussed in Section 2.4.

Stages may impact on each other in terms of the level of trust they are able to propagate. Romei et al. [77] reviewed various cases studies that led to biased decisions and analysed the causes. There is an important specific case of this, namely that trust impact may only manifest itself in later stages or at a later time. For instance, in the above example of improving the 'Data Preparation', this enhancement will only impact trust by users if it is made visible to these users. This may for instance be through better results when using the services, but, possibly more likely, may also only become visible if news article or long-term statistics are presented to users that explain that results are, say, less biased and that therefore can be trusted. However, where trust is established or damaged based on visible outcomes at later stages the resulting levels of trust will have implications for trust in all stages of the development of future technologies.

The second dimension present in the Chain of Trust results from the fact that a service or product may iterate through the stages during its lifecycle, possibly multiple times. This occurs, for instance, when new data is being introduced to improve the 'Training', and through this, the 'Inference'. In this case, effectively the service cycles through the chain depicted in Figure 2.

A third dimension within the notion of Chain of Trust is the development of trust through the stages of design, development and deployment of the AI-based service or product. Trust will be impacted by technology decisions in all stages of the lifecycle. In the above examples, we mainly consider the deployment stage, in which trust is considered from the perspective of a running service or existing product. Even simply making an AI-based service available may run the risk of introducing new biases, or exacerbating existing ones, because changing the way a service is offered may imply it is less useful or effective for certain groups. For instance, a recent study demonstrates inherent discrimination in automated mortgage advising in the FinTech industry [9]. Therefore, to establish trusted AI-based solutions it will be critical to consider trust from the initial stages, starting from the design of a new service or product. In so doing, trust is considered *a priori*, before it is being deployed, and does not come as a surprise once the service is running.

A final dimension to consider in the context of the Chain of Trust is that of accidents, sudden breakdowns of trust or failures. Typical examples of such trust failures are security breaches that impact trust, (reports about) data loss of the service or similar services, or the discovery of bias in the machine learning results that drive a service. Such accidents can take place through all pipeline and life cycle stages discussed above and may have severe impact on the level of trust users place in AI-based systems. However, the ways in which an organisation responds to such accidents can be equally, or more, important for determining the impact on trust. Dietz and Gillespie [30] have shown that scandals or crises which risk damaging the reputation of an organisation can also act as catalysts for culture change bringing about and reinforcing new ethical/trustworthy practice. They are opportunities to forge new relationships with stakeholders (positive or negative) [40]. Technology solutions that continuously monitor and possibly transparently share data about service bias are trustworthy technologies that may assist in avoiding or mitigating the impact of trust failures.

The Chain of Trust denotes the stages in which one can and should consider the trust qualities of machine learning technologies, both for initial trust and continuous trust. The chain of Trust also identifies the opportunities to maintain the trustworthiness of the system given the evolving nature of the relationship between the system and its users. Finally, the Chain of Trust provides guidance for experts and the public how and when to evaluate the trustworthiness of the system in relation to the ABI+ framework, not only in the final outcome of the system, but also in all the inner stages that leads to such outcome.

4 TRUST IN AI-BASED SYSTEMS

We now aim to establish the connection between *trust* in the AI-based solution and *trustworthiness* of the underlying technologies. We first discuss in Section 4.1 various trust related issues one may encounter in AI-based services and products. We then discuss in Section 4.2 a variety of existing Principled AI policy and technology frameworks that have emerged in recent years. Based on these frameworks, we will propose in Section 4.3 a technology-inspired Principled AI variant, namely FEAS Technologies, that is, technologies for Fairness, Explainability, Auditability and Safety.

4.1 Trust Considerations in Stages of the Chain of Trust

In this section we discuss trustworthiness aspects in the various stages identified in the Chain of Trust. We divide the discussion in data-related concerns (with the focus on data collection, data pre–processing and feature selection, Section 4.1.1) and model-related concerns (with the focus on the stages of model training, testing, and inference, Section 4.1.2).

4.1.1 Data Related Trust Concerns: Data Collection and Pre-Processing Stages. Data collection involves reading data from various sources reliably (e.g. sensors to collect environmental data, a smart speaker listening to audio commands or website that stores session-related data from the user's browser, or etc.); secure transmission of the data from the collection point to a machine for the purpose of storage or online analysis; storing data in server. The pre–processing stage includes mechanisms to clean the dataset by removing null, duplicate or noisy data. Usually, pre–processing solutions combines with the other ones in data–related stages.

One of the main trust concerns in data is in protection against privacy violations and other forms of illegitimate use of one's data. Legislative support (e.g, in Europe via GDPR regulation) has an important role to play. GDPR is a perfect illustration of several data-related trust issues. For instance, the collection process should be transparent and requires explicit consent from users for many

data uses. Moreover, the subject keeps the right to challenge the ability to collect data, has control over their collected data and maintains the "right to be forgotten". GDPR-type concerns and solutions provides a basis to identify technologies that enhance trust.

In general, the perceived intrusive nature of instances of data collection has convinced common users to be more cautious in the situations that their data is being collected [22]. This is a threat to trustworthiness of a service that functions based on collected data, particularly AI-based services and products.

Technological solutions for trustworthy data collection, preprocessing as well as storage have been well established. Trustworthiness usually relies on factual declaration of the good faith (benevolence) and abiding to it in action (integrity). The current trust challenges are therefore less in the development of technology solutions than in identifying ways of interacting, working and regulating aspects that impact trust.

4.1.2 Model-Related Trust Concerns: Feature Extraction, Training, Testing and Inference Stages. Model-related trust concerns trust in the working of the models and algorithms. This set of concerns gets to the heart of trust challenges we phase in a world in which AI becomes omni-present, fundamentally changing the way people live their lives. Many of the trust concerns relate to FAT, a fear that algorithms may harm society, or individuals, because they are unfair, without accountability, and non-transparent. As for data (Section 4.1.1) GDPR is useful as an illustration of the issues important for society. GDPR enforces a requirement to "explain" results of AI-based solutions to end users, which implies a desire for more transparent machine learning. GDPR also contains substantial accountability measures to implement a mechanism to "challenge" the outcome of AI–based technology.

The question is what the role of technologies in dealing with model-related trust concerns. At their core, machine learning models and algorithms are optimised for accuracy of results and efficiency in obtaining these. While the accuracy and speed of results might be considered to demonstrate basic functionality (ability in ABI+ terminology, Section 2.1), they do not necessarily satisfy or align with other trust qualities. In most cases, algorithms are considered as a "black-box" [68], which implies algorithms can be assessed only in relation to the outcomes they produce. Assessments of the benevolence or integrity (again using ABI+ terms) can therefore only be done in indirect manners, through that of the entity that develops or applies the AI-based solution. Of course, justifying such trust is problematic, especially in the time of high profile data breaches and scandals relating to mishandling or misuse of personal data (e.g. Facebook, Cambridge Analytica).

In Section 5 we will introduce a number of technologies to enhance or impact trust, of two types. The first type is to establish a mechanism to verify the outcomes of the model. In this case, an agent would be responsible to function in parallel to the model and the model outcome are not accessible until the agent and the model are both satisfied in a pre–defined criteria. In the second type one endeavours to (re)design a model or choice of algorithms into something that is inherently more trustworthy. For example, the design of a fair SVM model refers to embedding a fairness constraints into its definition so that the model functions with the built-in consideration of that notion. This set of approaches we will discuss in detail in Section 5.

4.2 Principled AI Policy Frameworks

The trust concerns discussed in the previous section are related to concerns that have been raised widely about the impact on society of the proliferation of AI. This has resulted in the emergence of a large amount of policy frameworks that relate to Principled AI frameworks, that is, policy frameworks to enhance and regulate Fairness, Accountability and Transparency of, particularly, AI-based services and products. Principled AI frameworks are particularly relevant to trust as well. Therefore, as depicted in Figure 1, the bottom box, it is opportune to relate Principled AI frameworks to the trustworthy technology classification we introduce in Section 4.3. It is important to note that in this paper we use FEAS to classify *technologies*, while many of the Principled AI frameworks inform *policy* and do not provide much detail in terms of specifying or restricting technology implementations to achieve the policy objectives.

Principled AI frameworks have been introduced by various stakeholders (technology companies, professional, standardisation, governmental and legislator bodies, academic researchers), as illustrated by Table 1. These Principled AI frameworks present varying sets of qualities that AI-based systems should follow to be considered trustworthy (some Principled AI frameworks (also) apply to technologies other than AI). In general, these documents present high–level definitions for the objectives and qualities of the involved science and technology, but do not go into the specifics of technical implementation guidelines.

There is emerging literature reviewing Principled AI frameworks. Whittlestone et al. [88] provide a critical analysis of frameworks for ethical machine learning and highlights a number of challenges, some of which require attention from a technical perspective. For instance, frameworks may confuse the interpretation of qualities, present conflicting definitions and/or qualities may be different across different documents. Particularly relevant also for the underlying technologies is that the frameworks often fail to realise dependencies between policy objectives (e.g. addressing discrimination issues might lead to unexpected privacy leakages). Current frameworks are focused on privacy, transparency and fairness issues but this needs to be shifted toward understanding such tensions and re-framing core research questions.

In yet unpublished work, Fjeld et al. [38] analyse currently available Principled AI frameworks, from industry, governments, general public, civil societies and academic bodies. Table 1 contains many of the Principled AI frameworks considered by [38] (see Section 4.3 for an explanation of how we compiled Table 1). Interestingly, they recognised 47 qualities, categorised them into eight groups, which are a combination of the qualities identified by Siau [59], i.e., humane (promotion of human values, professional responsibility, human control of technology) and technological qualities (fairness and non–discrimination, transparency and explainability, safety and security, accountability, privacy). The authors have available a graphical demonstration of their findings [1].

[1] https://ai-hr.cyber.harvard.edu/primp-viz.html

We note again that most of the frameworks do not focus on trust, but on ethics, privacy and related concerns. Moreover, the terms *ethical* and *trustworthy* machine learning are at times used interchangeably in these frameworks [52]. This would effectively imply that trustworthiness is achieved through abiding with the ethical concepts such as human rights or non-discrimination approaches. However, while ethical considerations are inevitably related to perceptions of trust, ethical machine learning and trustworthy machine learning are not necessarily the same thing. Specifically, in terms of ABI+, ethical machine learning would necessarily emphasise the benevolence aspects of trust, while the other two aspects critical for trust (ability and integrity) are insufficiently represented.

4.3 FEAS: Fair, Explainable, Auditable and Safe Technologies

We propose to classify trustworthy technologies in Fair, Explainable, Auditable and Safe Technologies (FEAS). This is in part motivated by a desire to align our discussion of trustworthy technologies with the Principled AI frameworks that are available in the literature, as discussed in Section 4.2. We stress that the implementation of such technological solutions in itself will not make the system trusted. Clearly, trust is only to a limited extend a technology challenge, which is the reason we provide in this paper the linkage of technologies with non-technological perspectives on trust. Moreover, even if one considers only technology, FEAS technologies are not the only ones that impact trust. For instance, a user–friendly Graphical User Interface or thoughtful presentation of the results through meaningful plots may have an impact on the overall trust in the system, too. However, FEAS technologies represent technological advances focused specifically on machine-learning and are therefore the focus of our attention. The question remains to what extent the implementation of FEAS technologies, which aim to enhance trustworthiness, actually also enhance people's trust. This is an open problem and should be investigated thoroughly in the future. We converged on FEAS based on our knowledge and understanding of the technologies involved, classified in manner we believe will be comprehensible, illustrative and natural for technologists. Note that the FEAS technological qualities are in addition to the essential technological qualities of *accuracy* and *efficiency* and performance of the algorithm(s), without which trustworthiness is not possible.

- **Fairness Technologies:** technologies focused on detection or prevention of discrimination and bias in different demographics [23, 25, 35, 37, 46, 49, 50, 55, 63, 77, 96–98].
- **Explainability Technologies:** technologies focused on explaining and interpreting the outcome to the stakeholders (including end-users) in a humane manner [6, 17, 19, 33, 41, 42, 60, 62, 69, 76, 86, 89].
- **Auditability Technologies:** technologies focused on enabling third-parties and regulators to supervise, challenge or monitor the operation of the model(s) [2, 8, 10, 18, 20, 21, 27, 28, 66, 74, 90, 94].
- **Safety Technologies:** technologies focused on ensuring the operation of the model as intended in presence of active or passive malicious attacker [14–16, 56, 71, 73, 78, 79].

4.3.1 FEAS Related to Principled AI. . Table 1 provides the relationship between the FEAS technology classes and the Principled AI frameworks identified in [38]. We reviewed each of the frameworks with respect to the FEAS technology classes required to establish the qualities mentioned in the framework. We mark frameworks that refer to fairness, explainability, safety and auditability qualities using the symbols explained in the caption of Table 1.

As one sees immediately from Table 1 all frameworks are related to FEAS technologies, and would be able to make use, or even require, FEAS technologies to be available. Note that there is a considerable difference in the granularity of the discussions in the Principled AI frameworks compared to that of the computing literature. Hence, the precise technology needs for each framework would need deeper investigations, and may not be completely specified within the existing framework documents. For instance, the policy frameworks refer to the general existence of discrimination caused by bias in machine learning algorithms, but in the technological literature there are at least 21 mathematical definitions for fairness and a wide range of solutions to prevent/detect bias in the algorithm. The technology discussion in Section 5 is therefore at a much deeper level of detail than that of the Principled AI frameworks of Table 1.

4.3.2 FEAS Related to Trust Qualities. Table 2 provides the relationship between trust qualities (humane, environmental and technological, see Section 2.3) and FEAS technologies. Table 2 is based on the authors' understanding of the qualities and technologies, the latter to be discussed deeper in Section 5. Fairness strongly requires technologies that are strong in humane and environmental qualities, as does explainability, since their effectiveness strongly depends on individuals and the culture or setting. Safety is dominated by technological quality associated with security and reliability of the systems.

5 TRUSTWORTHY MACHINE LEARNING TECHNOLOGIES

This section discuss technologies for trustworthy machine learning using the FEAS grouping introduced in the previous section. We introduce the FEAS classes, discuss challenges and provide some examples of existing approaches. A full review of technologies is beyond the scope of this paper.

5.1 Fairness Technologies

This group of technologies is concerned about achieving fair, non-discriminating outcomes. The ethical aspects of fairness constitute a structural assurance in the service or product that enhances (or at least impacts) trust.

Fair machine learning is a difficult challenge. A first challenge is to identify if how to measure unfairness, typically in terms of bias or related notions. Narayanan [70] has identified at least 21 definitions of fairness in the literature, which cannot necessarily all be obtained at the same time. To enhance trust, the metrics used by machine learning experts needs to relate to how it impacts trust by individuals and the public, posing an additional challenge.

Various solutions have been proposed to establish a subset of fairness notions. One approach is to reduce the bias in the dataset,

Table 1: Trustworthy technology classes related to FAT* frameworks. ✗= *no mention*, ✓= *mentioned*, ✔✔= *emphasised*

Framework	Year	Document Owner	Entities	Country	Fairness	Explainability	Safety	Auditability
Top 10 principles of ethical AI	2017	UNI Global Union	Ind	Switzerland	✓	✓	✓	✓
Toronto Declaration	2018	Amnesty International	Gov, Ind	Canada	✔✔	✓	✗	✓
Future of work and Education For the Digital Age	2018	T20: Think 20	Gov	Argentina	✔✔	✓	✓	✓
Universal Guidelines for AI	2018	The public voice coalition	Ind	Belgium	✔✔	✓	✓	✓
Human Rights in the Age of AI	2018	Access Now	Gov, Ind	United States	✔✔	✓	✔✔	✓
Preparing for the Future of AI	2016	US national Science, and Technology Council	Gov, Ind, Acad	United States	✓	✓	✔✔	✓
Draft AI R&D Guidelines	2017	Japan Government	Gov	Japan	✗	✓	✔✔	✓
White Paper on AI Standardization	2018	Standards Administration of China	Gov, Ind	China	✓	✗	✔✔	✔✔
Statements on AI, Robotics and 'Autonomous' Systems	2018	European Group on Ethics in Science and New Technologies	Gov, Ind, Acad	Belgium	✓	✓	✓	✔✔
For a Meaningful Artificial Intelligence	2018	Mission assigned by the French Prime Minister	Gov, Ind	France	✓	✓	✗	✔✔
AI at the Service of Citizens	2018	Agency for Digital Italy	Gov, Ind	Italy	✓	✓	✓	✓
AI for Europe	2018	European Commission	Gov, Ind	Belgium	✓	✓	✓	✓
AI in the UK	2018	UK House of Lords	Gov, Ind	United Kingdom	✔✔	✓	✓	✓
AI in Mexico	2018	British Embassy in Mexico City	Gov	Mexico	✓	✗	✓	✓
Artificial Intelligence Strategy	2018	German Federal Ministries of Education, Economic Affairs, and Labour and Social Affairs	Gov, Ind	Germany	✓	✓	✓	✔✔
Draft Ethics Guidelines for Trustworthy AI	2018	European High Level Expert Group on AI	Gov, Ind, Civ		✔✔	✓	✔✔	✓
AI Principles and Ethics	2019	Smart Dubai	Ind	UAE	✔✔	✓	✔✔	✓
Principles to Promote FEAT AI in the Financial Sector	2019	Monetary Authority of Singapore	Gov, Ind	Singapore	✓	✓	✗	✓
Tenets	2016	Partnership on AI	Gov, Ind, Acad	United States	✓	✓	✓	✓
Asilomar AI Principles	2017	Future of Life Institute	Ind	United States	✓	✓	✓	✓
The GNI Principles	2017	Global Network Initiative	Gov, Ind	United States	✗	✓	✓	✓
Montreal Declaration	2018	University of Montreal	Gov, Ind, Civ	Canada	✔✔	✔✔	✔✔	✓
Ethically Aligned Design	2019	IEEE	Ind	United States	✓	✓	✓	✔✔
Seeking Ground Rules for AI	2019	New York Times	Ind, GeP	United States	✓	✓	✓	✓
European Ethical Charter on the Use of AI in Judicial Systems	2018	Council of Europe: CEPEJ	Gov	France	✓	✓	✓	✓
AI Policy Principles	2017	ITI	Gov, Ind	United States	✓	✓	✔✔	✓
The Ethics of Code	2017	Sage	Ind	United States	✓	✗	✗	✓
Microsoft AI Principles	2018	Microsoft	Ind	United States	✓	✓	✔✔	✓
AI at Google: Our Principles	2018	Google	Ind	United States	✓	✓	✔✔	✓
AI Principles of Telef\'onica	2018	Telef\'onica	Ind	Spain	✓	✓	✓	✗
Guiding Principles on Trusted AI Ethics	2019	Telia Company	Ind	Sweden	✓	✓	✔✔	✓
Declaration of the Ethical Principles for AI	2019	IA Latam	Ind	Chile	✓	✓	✔✔	✓

Table 2: Trustworthy technology classes versus trust qualities [59]. ✗= *less important*, ✓= *important*, ✔✔*very important*

	Fair	Explainable	Auditable	Safe
Humane Qualities	✔✔	✔✔	✓	✗
Technological Qualities	✓	✓	✗	✔✔
Environmental Qualities	✔✔	✔✔	✓	✓

known as *debiasing data*. However, it is not sufficient (or even helpful) to simply ignore or remove features associated with unfairness, e.g. gender, ethnics [63]. Luong et al. [65] assigned a decision value to each data sample and adjusted this value to eliminate discrimination. Kamishima et al. [50] proposed a model–specific approach by adding a regulating term for a logistic regression classifier, which eventually leads to unbiased classification. Other fairness mitigation solutions focus on the inference stage. They enforce the output of the model to generate a specific notion of fairness [46].

Žliobaitė et al. [98] identify two conditions for non-discriminating machine learning: data with the same non-protected attributes

should give the same outcomes, and the ability to distinguish outcomes should be of the same order as the difference in the non-protected attribute values. This provides a more generic understanding helpful to design fairness technologies. This paper does not aim to review all techniques but it is clear that many challenges remain in achieving fair machine learning technologies.

5.2 Explainability Technologies

Explainability refers to relating the operation and outcomes of the model into understandable terms to a human [51]. In the machine learning literature, notions of *explainability*, *transparency*, *intelligibility*,*comprehensibility* and *interpretability* are often used interchangeably [43]. Lipton [60] provides a thorough discussion on these terms and the differences. He also concludes that explainability increases trust. Doshi-Velez et al. [31] suggested two types of explainability, namely explainability of an application (e.g. a physician being able to understand the reasons for a classifier's medical diagnostic [75]) and explainability to understand the way in which a classifier is coming up with its outputs, mainly by using intelligible models.

There are two main approaches in explainable solutions. The first one, known as *ex-ante*, refers to the use of highly intelligible models to obtain the desired predictions. The second one, known as *ex-post*, refers to the use of a second model to understand the learned model's behaviour. In ex–ante approach, an explicit prediction function analyses feature coefficients to understand their impact over a decision, decision trees or decision lists [76]).

Ex-post explanations are categorised into *local* and *global* explainers. The ex-post approach uses explain. There is a fast growing body of literature, including local explainers (e.g., [75, 80] and the unified framework for local explainers which generalises these and other existing methods [64]), global explainers (e.g., [57]). There are many challenges left with respect to explainable machine learning technologies, including trading off fidelity to the original model and explainability [69].

5.3 Auditability Technologies

Auditability technologies refer to methods that enable third parties to challenge the operation and outcome of a model. This provides more transparency to black–box characteristics of machine learning algorithms. Enabling machine learning lineage provides insights into how they have operate, which gives a greater degree of transparency compared to explainability of current processes or outcomes.

More specifically, auditability in machine learning may involve assessing the influence of input data in the output of the model. This ensures predictability, a process that is also referred as "decision provenance" in the literature [82]. Singh et al. [82] argue that decision provenance have two objectives, it should provide the history of particular data and it should provide a viewpoint on the system's behaviour and its interactions with its inner components or outside entities.

Much of the literature around auditability relates to data provenance research in database and cloud concepts [44] and model provenance approaches [39]. It involves proposals on how to store provenance data [18, 20, 21, 74, 94], summarisation of provenance data [2, 28, 66], specific query language for the provenance data [28], query explainability [8, 10], Natural language processing for provenance data [27, 90] and cryptographic solutions to verify model's behaviours without exposing user privacy [72] or revealing model's intellectual properties [87].

5.4 Safety Technologies

Data and the machine learning model could both be the target to adversarial operations. The extent of attacker's access to the data and model depends on the intention of the attack and the weaknesses in the system's architecture. The attacker can execute a targeted attack to harm an individual or perform a indiscriminate attack. Moreover, the attacker can perform the attack stealthily for the intention of information gathering or intelligence (a.k.a exploratory attack) or he/she can actively engage into the functioning of the system for the purpose of manipulation (a.k.a causative attack).

The security and privacy foundations of a ML model is not different from classical security model of Confidentiality, Integrity and Availability (CIA model [73]). We omit availability, since its relevance is general, not just or specifically for AI-based services. Careless preparation of the stored data would leak information to an attacker [61] but confidentiality can be enhanced in many ways, for instance through approaches for differential privacy [1, 34], homomorphic encryption [47] or cryptography integrated in machine learning algorithms [79]. Integrity can be enhanced by either preventing tampering, such as in pre-processing [73], or by discovering and possibly repairing tampered data [12, 14, 58]. These methods deal only with data, but it is also relevant to consider integrity during the execution of the algorithm, e.g., in the training stage [11, 13, 26, 61, 85].

6 CONCLUSION

This paper established the connection between trust as a notion within the social sciences, and the set of technologies that are available for trustworthy machine learning. More specifically, we related the ABI+ framework and HET technology qualities for trust with categories of machine learning technologies that enhance trustworthiness. We identified four categories of technologies that need to be considered: Fair, Explainable, Auditable and Safe (FEAS) technologies. These need to be considered in various interrelated stages of a system life cycle, each stage forming part of a Chain of Trust. The paper shows a close relationship between technologies to improve the trustworthiness of AI-based systems and those that are being pursued in ethical AI and related endeavours. We illustrated this by mapping of FEAS technologies on concerns in a large set of international Principled AI policy and technology frameworks.

ACKNOWLEDGEMENTS

This work was funded in part by the UK Engineering and Physical Sciences Research Council for the projects titled "Fintrust: Trust Engineering for the Financial Industry" (EP/R033595/1) and "EPSRC Centre for Doctoral Training in Cloud Computing for Big Data" (EP/L015358/1).

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 308–318.

[2] Eleanor Ainy, Pierre Bourhis, Susan B Davidson, Daniel Deutch, and Tova Milo. 2015. Approximated summarization of data provenance. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 483–492.

[3] Mhairi Aitken, Sarah Cunningham-Burley, and Claudia Pagliari. 2016. Moving from trust to trustworthiness: Experiences of public engagement in the Scottish Health Informatics Programme. *Science and Public Policy* 43, 5 (2016), 713–723.

[4] Naomi S Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46, 3 (1992), 175–185.

[5] Peter Andras, Lukas Esterle, Michael Guckert, The Anh Han, Peter R Lewis, Kristina Milanovic, Terry Payne, Cedric Perret, Jeremy Pitt, Simon T Powers, and others. 2018. Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems. *IEEE Technology and Society Magazine* 37, 4 (2018), 76–83.

[6] Susan Athey and Guido W Imbens. 2015. Machine learning methods for estimating heterogeneous causal effects. *stat* 1050, 5 (2015), 1–26.

[7] A Avizienis, J . Laprie, B Randell, and C Landwehr. 2004. Basic concepts and taxonomy of dependable and secure computing. *IEEE Transactions on Dependable and Secure Computing* 1, 1 (1 2004), 11–33. https://doi.org/10.1109/TDSC.2004.2

[8] Akanksha Baid, Wentao Wu, Chong Sun, AnHai Doan, and Jeffrey F Naughton. 2015. On Debugging Non-Answers in Keyword Search Systems.. In *EDBT*. 37–48.

[9] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. 2018. Consumer-Lending Discrimination in the Era of FinTech. *Unpublished working paper. University of California, Berkeley* (2018).

[10] Nicole Bidoit, Melanie Herschel, and Katerina Tzompanaki. 2014. Query-based why-not provenance with nedexplain. In *Extending database technology (EDBT)*.

[11] Battista Biggio, Igino Corona, Giorgio Fumera, Giorgio Giacinto, and Fabio Roli. 2011. Bagging classifiers for fighting poisoning attacks in adversarial classification tasks. In *International workshop on multiple classifier systems*. Springer, 350–359.

[12] Battista Biggio, Igino Corona, Blaine Nelson, Benjamin I P Rubinstein, Davide Maiorca, Giorgio Fumera, Giorgio Giacinto, and Fabio Roli. 2014. Security evaluation of support vector machines in adversarial environments. In *Support Vector Machines Applications*. Springer, 105–153.

[13] Battista Biggio, Giorgio Fumera, and Fabio Roli. 2010. Multiple classifier systems for robust classifier design in adversarial environments. *International Journal of Machine Learning and Cybernetics* 1, 1-4 (2010), 27–41.

[14] Battista Biggio, Giorgio Fumera, and Fabio Roli. 2014. Security evaluation of pattern classifiers under attack. *IEEE transactions on knowledge and data engineering* 26, 4 (2014), 984–996.

[15] Battista Biggio, Blaine Nelson, and Pavel Laskov. 2011. Support vector machines under adversarial label noise. In *Asian Conference on Machine Learning*. 97–112.

[16] Matt Bishop. 2007. About penetration testing. *IEEE Security & Privacy* 5, 6 (2007), 84–87.

[17] Avrim L Blum and Pat Langley. 1997. Selection of relevant features and examples in machine learning. *Artificial Intelligence* 97, 1-2 (1997), 245–271.

[18] Peter Buneman, Adriane Chapman, and James Cheney. 2006. Provenance management in curated databases. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*. ACM, 539–550.

[19] Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. 2018. Feature selection in machine learning: A new perspective. *Neurocomputing* 300 (2018), 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

[20] Adriane P Chapman, Hosagrahar V Jagadish, and Prakash Ramanan. 2008. Efficient provenance storage. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 993–1006.

[21] James Cheney, Amal Ahmed, and Umut A Acar. 2007. Provenance as dependency analysis. In *International Symposium on Database Programming Languages*. Springer, 138–152.

[22] Erika Chin, Adrienne Porter Felt, Vyas Sekar, and David Wagner. 2012. Measuring user confidence in smartphone security and privacy. In *Proceedings of the eighth symposium on usable privacy and security*. ACM, 1.

[23] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv preprint arXiv:1808.00023* (2018).

[24] Sarah Cunningham-Burley. 2006. Public knowledge and public trust. *Public Health Genomics* 9, 3 (2006), 204–210.

[25] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. 2017. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data* 5, 2 (2017), 120–134.

[26] Ambra Demontis, Marco Melis, Battista Biggio, Davide Maiorca, Daniel Arp, Konrad Rieck, Igino Corona, Giorgio Giacinto, and Fabio Roli. 2017. Yes, machine learning can be more secure! a case study on android malware detection. *IEEE Transactions on Dependable and Secure Computing* (2017).

[27] Daniel Deutch, Nave Frost, and Amir Gilad. 2016. Nlprov: Natural language provenance. *Proceedings of the VLDB Endowment* 9, 13 (2016), 1537–1540.

[28] Daniel Deutch, Amir Gilad, and Yuval Moskovitch. 2015. Selective provenance for datalog programs using top-k queries. *Proceedings of the VLDB Endowment* 8, 12 (2015), 1394–1405.

[29] Graham Dietz and Deanne N Den Hartog. 2006. Measuring trust inside organisations. *Personnel review* 35, 5 (2006), 557–588.

[30] Graham Dietz and Nicole Gillespie. 2012. *Recovery of Trust: Case Studies of Organisational Failures and Trust Repair*. Vol. 5. Institute of Business Ethics London.

[31] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[32] Harris Drucker, Donghui Wu, and Vladimir N Vapnik. 1999. Support vector machines for spam categorization. *IEEE Transactions on Neural networks* 10, 5 (1999), 1048–1054.

[33] Rehab Duwairi and Mahmoud El-Orfali. 2014. A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. *Journal of Information Science* 40, 4 (2014), 501–513.

[34] Cynthia Dwork. 2011. Differential privacy. *Encyclopedia of Cryptography and Security* (2011), 338–340.

[35] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 214–226.

[36] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. 2010. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research* 11, Feb (2010), 625–660.

[37] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.

[38] Jessica Fjeld, Hannah Hilligoss, Nele Achten, Maia Levy Daniel, Sally Kagay, and Joshua Feldman. 2019. Principled Artificial Intelligence: Mapping Consensus and Divergence in Ethical and Rights-Based Approaches. (2019). https://ai-hr.cyber.harvard.edu/

[39] Zahra Ghodsi, Tianyu Gu, and Siddharth Garg. 2017. Safetynets: Verifiable execution of deep neural networks on an untrusted cloud. In *Advances in Neural Information Processing Systems*. 4672–4681.

[40] Nicole Gillespie and Graham Dietz. 2009. Trust repair after an organization-level failure. *Academy of Management Review* 34, 1 (2009), 127–145.

[41] Carlos Adriano Gonçalves, Celia Talma Gonçalves, Rui Camacho, and Eugenio C Oliveira. 2010. The impact of Pre-Processing on the Classification of MEDLINE Documents. *Pattern Recognition in Information Systems, Proceedings of the 10th International Workshop on Pattern Recognition in Information Systems, PRIS 2010, In conjunction with ICEIS 2010* (2010), 10.

[42] Carlos Vladimiro González Zelaya. 2019. Towards Explaining the Effects of Data Preprocessing on Machine Learning. *2019 IEEE 35th International Conference on Data Engineering (ICDE)* (2019).

[43] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 93.

[44] Himanshu Gupta, Sameep Mehta, Sandeep Hans, Bapi Chatterjee, Pranay Lohia, and C Rajmohan. 2017. Provenance in context of Hadoop as a Service (HaaS)-State of the Art and Research Directions. In *2017 IEEE 10th International Conference on Cloud Computing (CLOUD)*. IEEE, 680–683.

[45] Rob Hagendijk and Alan Irwin. 2006. Public deliberation and governance: engaging with science and technology in contemporary Europe. *Minerva* 44, 2 (2006), 167–184.

[46] Moritz Hardt, Eric Price, Nati Srebro, and others. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[47] Weiwei Hu and Ying Tan. 2017. Generating adversarial malware examples for black-box attacks based on GAN. *arXiv preprint arXiv:1702.05983* (2017).

[48] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. 1999. Data clustering: a review. *ACM computing surveys (CSUR)* 31, 3 (1999), 264–323.

[49] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2016. Rawlsian Fairness for Machine Learning. *FATML* (2016), 1–26. http://arxiv.org/abs/1610.09559

[50] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[51] Been Kim, Elena Glassman, Brittney Johnson, and Julie Shah. 2015. iBCM: Interactive Bayesian case model empowering humans via intuitive interaction. (2015).

[52] Sabine Theresia Koszegi. 2019. High-Level Expert Group on Artificial Intelligence.

[53] Alex Krizhevsky, Geoffrey Hinton, and others. 2009. *Learning multiple layers of features from tiny images*. Technical Report. Citeseer.

[54] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[55] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

[56] Ricky Laishram and Vir Virander Phoha. 2016. Curie: A method for protecting SVM Classifier from Poisoning Attack. *arXiv preprint arXiv:1606.01584* (2016).

[57] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & Explorable Approximations of Black Box Models. (7 2017).

[58] Pavel Laskov and Marius Kloft. 2009. A framework for quantitative security analysis of machine learning. In *Proceedings of the 2nd ACM workshop on Security and artificial intelligence*. ACM, 1–4.

[59] Xin Li, Traci J Hess, and Joseph S Valacich. 2008. Why do we trust new technology? A study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems* 17, 1 (2008), 39–71.

[60] Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).

[61] Qiang Liu, Pan Li, Wentao Zhao, Wei Cai, Shui Yu, and Victor C M Leung. 2018. A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE access* 6 (2018), 12103–12117.

[62] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13* (2013), 623. https://doi.org/10.1145/2487575.2487579

[63] Kristian Lum and James Johndrow. 2016. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077* (2016).

[64] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.

[65] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 502–510.

[66] Peter Macko, Daniel Margo, and Margo Seltzer. 2013. Local clustering in provenance graphs. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 835–840.

[67] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[68] Donald Michie, David J Spiegelhalter, C C Taylor, and others. 1994. Machine learning. *Neural and Statistical Classification* 13 (1994).

[69] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73 (2018), 1–15.

[70] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*.

[71] Olga Ohrimenko, Felix Schuster, Cédric Fournet, Aastha Mehta, Sebastian Nowozin, Kapil Vaswani, and Manuel Costa. 2016. Oblivious multi-party machine learning on trusted processors. In *25th ${USENIX$}$ Security Symposium (${USENIX$}$ Security 16)*. 619–636.

[72] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. 2016. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814* (2016).

[73] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. 2018. SoK: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 399–414.

[74] Christopher Ré and Dan Suciu. 2008. Approximate lineage for probabilistic databases. *Proceedings of the VLDB Endowment* 1, 1 (2008), 797–808.

[75] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*. ACM, 1135–1144.

[76] Ronald L Rivest. 1987. Learning decision lists. *Machine learning* 2, 3 (1987), 229–246.

[77] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29, 5 (2014), 582–638.

[78] Benjamin I P Rubinstein, Blaine Nelson, Ling Huang, Anthony D Joseph, Shing-hon Lau, Satish Rao, Nina Taft, and J Doug Tygar. 2009. Antidote: understanding and defending against poisoning of anomaly detectors. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*. ACM, 1–14.

[79] Reza Shokri and Vitaly Shmatikov. 2015. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. ACM, 1310–1321.

[80] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 3145–3153.

[81] Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal* 31, 2 (2018), 47–53.

[82] Jatinder Singh, Jennifer Cobbe, and Chris Norval. 2018. Decision Provenance: Harnessing data flow for accountable systems. *IEEE Access* 7 (2018), 6562–6574.

[83] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[84] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[85] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[86] Alper Kursat Uysal and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Information Processing and Management* 50, 1 (2014), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006

[87] Michael Walfish and Andrew J Blumberg. 2015. Verifying computations without reexecuting them. *Commun. ACM* 58, 2 (2015), 74–84.

[88] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In *Proceedings of the AAAI/ACM Conference on AI Ethics and Society, Honolulu, HI, USA*. 27–28.

[89] Simon N Wood. 2003. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65, 1 (2003), 95–114.

[90] Eugene Wu, Samuel Madden, and Michael Stonebraker. 2013. Subzero: a fine-grained lineage system for scientific databases. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*. IEEE, 865–876.

[91] Brian Wynne. 1992. Misunderstood misunderstanding: social identities and public uptake of science. *Public understanding of science* 1, 3 (1992), 281–304.

[92] Brian Wynne. 1996. A reflexive view of the expert-lay knowledge divide. *Risk, environment and modernity: Towards a new ecology* 40 (1996), 44.

[93] Brian Wynne. 2006. Public engagement as a means of restoring public trust in science–hitting the notes, but missing the music? *Public Health Genomics* 9, 3 (2006), 211–220.

[94] Yulai Xie, Kiran-Kumar Muniswamy-Reddy, Dan Feng, Yan Li, and Darrell D E Long. 2013. Evaluation of a hybrid approach for efficient provenance storage. *ACM Transactions on Storage (TOS)* 9, 4 (2013), 14.

[95] Masaki Yuki, William W Maddux, Marilynn B Brewer, and Kosuke Takemura. 2005. Cross-cultural differences in relationship-and group-based trust. *Personality and Social Psychology Bulletin* 31, 1 (2005), 48–62.

[96] Muhammad Bilal Zafar. 2019. Discrimination in Algorithmic Decision Making: From Principles to Measures and Mechanisms. (2019).

[97] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1171–1180.

[98] Indrė Žliobait\.e. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* 31, 4 (2017), 1060–1089.