Towards Algorithm Auditing

*A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms**

Adriano Koshiyama[1,2,$], Emre Kazim[1,2,$], Philip Treleaven[1,$], Pete Rai[3], Lukasz Szpruch[4,5], Giles Pavey[1,6,7], Ghazi Ahamat[8], Franziska Leutner[1,9,10], Randy Goebel[11], Andrew Knight[12], Janet Adams[13], Christina Hitrova[14], Jeremy Barnett[1,15,16], Parashkev Nachev[1], David Barber[1], Tomas Chamorro-Premuzic[1,17,18], Konstantin Klemmer[19], Miro Gregorovic[20], Shakeel Khan[21,22], and Elizabeth Lomas[1]

[1]University College London, [2]Holistic AI, [3]Cisco Systems, [4]University of Edinburgh, [5]The Alan Turing Institute, [6]Unilever, [7]University of Oxford, [8]Centre For Data Ethics and Innovation, [9]HireVue, [10]Goldsmiths University of London, [11]University of Alberta, [12]Royal Institution of Chartered Surveyors, [13]Ainstein AI ™, [14]Technical University of Munich, [15]St Pauls Chambers, [16]Resilience Partners, [17]Columbia University, [18]ManpowerGroup, [19]University of Warwick, [20]London Stock Exchange, [21]UK HMRC, [22]ValidateAI

Abstract

Business reliance on algorithms are becoming ubiquitous, and companies are increasingly concerned about their algorithms causing major financial or reputational damage. High-profile cases include VW's Dieselgate scandal with fines worth of $34.69B, Knight Capital's bankruptcy (~$450M) by a glitch in its algorithmic trading system, and Amazon's AI recruiting tool being scrapped after showing bias against women. In response, governments are legislating and imposing bans, regulators fining companies, and the Judiciary discussing potentially making algorithms artificial "persons" in Law.

Soon there will be 'billions' of algorithms making decisions with minimal human intervention; from autonomous vehicles and finance, to medical treatment, employment, and legal decisions. Indeed, scaling to problems beyond the human is a major point of using such algorithms in the first place. As with *Financial Audit*, governments, business and society will require *Algorithm Audit*; formal assurance that algorithms are legal, ethical and safe. A new industry is envisaged: Auditing and Assurance of Algorithms (cf. Data privacy), with the remit to professionalize and industrialize AI, ML and associated algorithms.

The stakeholders range from those working on policy and regulation, to industry practitioners and developers. We also anticipate the nature and scope of the auditing levels and framework presented will inform those interested in systems of governance and compliance to regulation/standards. Our goal in this paper is to survey the key areas necessary to perform auditing and assurance, and instigate the debate in this novel area of research and practice.

[$]Corresponding authors: [adriano.koshiyama.15, e.kazim, p.treleaven]@ucl.ac.uk

*Listed authors are those who contributed substantive ideas and/or work to this report. Contributions include writing, research, and/or review for one or more sections, and/or via ongoing discussions. As such, except for the corresponding authors, inclusion as author does not imply endorsement of all aspects of the report; the other authors are listed roughly in order of contribution. The list of authors may change, subject to approval from their respective organizations.

1. Introduction

With the rise of Artificial Intelligence (AI), legal, ethical and safety implications of its use are becoming increasingly pivotal in business and society. We are currently entering a new phase of the 'digital revolution' in which privacy, accountability, fairness, and safety are becoming priority research and debate agendas for engineering and the social sciences (Treleaven et al., 2019; Brundage et al., 2020).

Like the 'Big Data' wave, this new phase of algorithmic decision making and evaluation ('Big Algo') can be paraphrased using the 5V's methodology:

- **Volume**: as resources and know-how proliferate, soon there will be 'billions' of algorithms;
- **Velocity**: algorithms making real-time decision with minimal human intervention;
- **Variety**: from autonomous vehicles to medical treatment, employment, finance, etc.;
- **Veracity**: reliability, legality, fairness, accuracy, and regulatory compliance as critical features;
- **Value**: new services, sources of revenue, cost-savings, and industries will be established.

Whilst the last decade (the 10s) the focus was on 'Data Protection', the shift now is towards 'Algorithm Conduct'. As a result, new technologies, procedures and standards will be needed to ensure that 'Big Algo' is an opportunity and not a threat to governments, business and society.

Algorithm Auditing is the research and practice of assessing, mitigating, and assuring an algorithm's safety, legality, and ethics. This area encompasses current research in areas such as AI Fairness, Explainability, Robustness, Privacy, as well as matured topics of Data ethics, management and stewardship. As with Financial Audit, eventually governments, business and society will require Algorithm Audit, i.e., the formal assurance that algorithms are legal, ethical and safe. In a snapshot, Figure 1 outlines the dimensions and examples of activities that are part of Algorithm Auditing. We define each one below.

- **Development**: the process of developing and documenting an algorithmic system.
- **Assessment**: the process of evaluating the algorithm behaviour and capacities.
- **Mitigation**: the process of servicing or improving an algorithm outcome.
- **Assurance**: the process of declaring that a system conforms to predetermined standards, practices or regulations.

A new industry is envisaged: Auditing and Assurance of Algorithms and Data, with the remit to professionalize and industrialize AI, ML and associated algorithms. Our goal with this paper is to instigate the debate in this novel area of Algorithm Audit. The following sections present the key components that cover the Algorithm Audit research and practice, namely: algorithms, verticals of auditing, mitigation strategy and assurance.

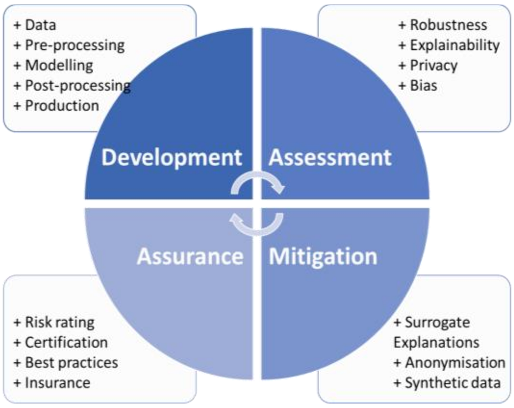


Figure 1. Dimensions and examples of activities that are part of algorithm auditing.

2. Key Components of Algorithm Auditing

In this section we describe the key parts encompassing Algorithm Auditing, namely: the algorithm as the centerpiece of the process; the main verticals of auditing; ways to perform auditing and what happens subsenquently; and finally, the possible outcome of auditing, namely, algorithm assurance processes.

2.1 Object of Audit: Algorithms

An algorithm is a finite sequence of well-defined, computer-implementable instructions, typically to solve a class of problems or to perform a computation. The key constituents of an algorithm are:

- **Data**: input, output, and simulation environment;
- **Model**: objective function, formulation, parameters and hyperparameters; and
- **Development**: design, documentation, building process and infra-structure, and open-source libraries.

In the 1980-1990s, Expert Systems (Giarratano and Riley, 1998) were mainly in vogue and the main concern in relation to quality assurance was restricted to **Development and Model** (Rushby, 1988). We should also mention that the focus during that period was more on accuracy and computational cost. Since the 00s, the paradigm has shifted, with now most of the industrial applications of AI relying on Machine Learning (Hastie et al., 2009; Sutton and Barto, 2018). This has added a new source of risk, namely **Data** (with model and data interacting in a much more complex way than before), to the quality assurance process; discussions are now broadly around bias and discrimination, interpretability and explainability, privacy, with a reduced focus on performance and resilience of early systems.

2.2 What to Audit: Verticals of Algorithm Auditing

Regardless of the algorithm, broadly speaking, there are five stages of Development (see Table 1):

I. **Data and Task Setup**: collecting, storing, extracting, normalising, transforming, and loading data. Ensuring that the data pipelines are well-structured, and the task (regression, classification, etc.) has been well-specified and designed. Ensuring that data and software artifacts are well documented and preserved.

II. **Feature pre-processing**: selecting, enriching, transforming, and engineering a feature space.

III. **Model selection**: running model cross-validation, optimization, and comparison.

IV. **Post-processing and Reporting**: adding thresholds, auxiliary tools and feedback mechanisms to improve interpretability, presenting the results to key stakeholders, evaluating the impact of the algorithmic system to the business.

V. **Productionizing and Deploying**: passing through several review processes, from IT to Business, and putting in place monitoring and delivery interfaces. Maintaining an appropriate record of in-field results and feedback.

Although these stages appear static and self-containing, in practice they interact in a dynamic fashion, not following a linear progression but a series of loops, particularly in between Pre/Post-processing.

Table 1. Interrelation between development stage and auditing verticals.

Stage	*Explainability*	*Robustness*	*Fairness*	*Privacy*
Data and Task Setup	Data collection and labelling	Data accuracy	Population balance	DPIA
Feature pre-processing	Dictionary of variables	Feature engineering	Fair representations	Data minimisation
Model selection	Model complexity	Model validation	Fairness constraints	Differential privacy
Post-processing and Reporting	Auxiliary tools	Adversarial testing	Bias metrics assessment	Model inversion
Productionizing and Deploying	Interface and documentation	Concept drift detection and continuous integration	Real-time monitoring of bias metrics	Rate-limiting and user's queries management

In Table 1 we also list how each stage interacts with four keys verticals:

- **Privacy**: quality of a system to mitigate personal or critical data leakage.
- **Fairness**: quality of a system to avoid unfair treatment of individuals or organizations.
- **Explainability**: quality of a system to provide decisions or suggestions that can be understood by their users and developers.
- **Robustness**: quality of a system to be safe, not vulnerable to tampering.

In a similar fashion to the stages, each vertical appears to be self-contained, but these are also interrelated. Though the research on each vertical is mostly conducted in silos, there is a growing reckoning from the scientific and industry community of the **Trade-offs and Interactions** between them. Accuracy, a component of robustness, may need to be traded for lowering any existing outcome metric of bias; making the model more explainable may affect some of the system performance and privacy; Improving privacy may affect ways to assess adverse impact of algorithmic systems; and so on. Optimisation of these features and tradeoffs will depend on multiple factors, notably the use case domain, the regulatory jurisdiction, and the risk appetite and values of the organization implementing the algorithm.

2.3 Ways to Audit: Levels of Access for Auditing

There are different levels of access that an auditor has during its investigation of an algorithm (see Table 3). In the scientific literature and technical reports, the commonplace is to categorize the knowledge about the system in two extremes: 'White-box' and 'Black-box'. In fact, the spectrum regarding the knowledge of a system is more of a continuum of 'shades of grey' than this simple dichotomy. This additional nuance allows for a richer exploration of the technologies available for assessment and mitigation, as well as the right level of disclosure that a certain business feels comfortable to engage.

Hence, we can identify seven levels of access that an auditor can have of a system. It ranges from the highest level 'White-box' where all the details encompassing the model are disclosed to the lowest level to 'Process-access' where only indirect observation of a system can be made. The levels in between are set by limiting the access to the components behind the learning process (e.g. knowledge of the objective function, model architecture, input data, etc.). Level 7 contains all the assessment, monitoring, and mitigation strategies of lower levels, with the report getting less detailed and inaccurate as levels decreases. Therefore, analysis and techniques requiring Level 7 cannot be used at Level 6 without proper assumptions and acceptable levels of inaccuracy.

2.4 After Audit: Mitigation Strategies

Given the feedback received as output of the audit interventions can be made to improve the systems outcome across the key verticals and stages. The more access to the algorithmic system, the more targeted, technical, diverse and effective will be the mitigation strategy employed. Table 2 lists possible interventions when a 'White-box' level is considered. When the access available is lower than 'White-box', then access to some stages and procedures are omitted from this table (e.g. Data and Task setup or Productionizing and Deploying).

Table 2. Interrelation between development stage and mitigation strategies for 'White-box' access level.

Stage	*Explainability*	*Robustness*	*Fairness*	*Privacy*
Data and Task Setup	Dictionary of variables and dataset sheets	Collecting targeted data, reframing loss function	Alternative data sources	Anonymisation
Feature pre-processing	Avoiding excessive feature engineering	Feature Squeezing	Synthetic data generation	Dimensionality Reduction
Model selection	By-design interpretable models	Adversarial Training	Counterfactual Fairness	Federated Learning
Post-processing and Reporting	LIME, SHAP	High Confidence Predictions and Confidence Intervals	Calibrated Odds	Model Inversion Mitigation
Productionizing and Deploying	Recourse interface	'Circuit-breaking'	Monitoring panels	Rate-limiting and user's queries management

2.5 Outcome of Audit: Assurance Processes

The broader outcome of an auditing process is to improve confidence or ensure trust of the underlying system and then to capture that in some certification process. After assessing the system, and implementing mitigation strategies, the auditing process assesses whether the system conforms to regulatory, governance and ethical standards. Providing assurance, therefore, needs to be understood through different dimensions and steps needs to be taken so that the algorithm can be shown to be trustworthy. Below we list key points that embody the assurance process:

- **General and sector-specific assurance:** broad national regulation and standards (provided agents such as NIST, UK-ICO, EU, etc.) with sector specific ones, such as in financial services (e.g. SEC, FCA, etc.), health (e.g. NIH, NHS, etc.), real estate (e.g. RICS, IVS, USPAP), etc.
- **Governance:** from two aspects, namely technical assessments (robustness, privacy, etc.) and impact (risk, compliance, etc.) assessments.
- **Unknown Risks:** discussing risk schemes and highlighting 'red teaming', which is used to mitigate unknown risks.
- **Monitoring Interfaces:** outlining risk assessments and the use of 'traffic-light' user friendly monitoring interfaces.
- **Certification:** the numerous ways in which certification may occur, such as certification of a system or AI engineers, etc.
- **Insurance**: a subsequent service to emerge as a result of assurance maturing.

Regulators face a growing challenge in both supervising the use of these algorithms amongst the sector which they oversee and the use of algorithms in their own regulatory process via RegTech and SupTech. There are some other 'soft' aspects, related to the governance structure underpinning the development. These are related to defining an algorithm's goals (what does it aim to achieve?), how it serves those it is making decisions about. These could compose a statement of intention whereby the designer sets out a position statement in advance indicating what it is that the algo is supposed to do. This could facilitate to judge whether the algo has performed as it was intended.

3. Algorithms

For completeness, this section unpacks algorithms across three domains: Computational Statistics (e.g. Monte Carlo methods), AI and ML (e.g. Artificial Neural Networks), and Complex Systems (e.g. Agent-Based systems). See below. While there may be some debate over the terminology, we find the classification helpful to distinguish between relatively well-established methods and more cutting-edge technologies.

- **Computational Statistics** - computationally intensive statistical methods.
- **Complex Systems** - systems with many interacting components whose aggregate activity is nonlinear and typically exhibit hierarchical self-organization under selective pressures.
- **AI Algorithms** - mimicking a form of learning, reasoning, knowledge, and decision-making
    - *Knowledge or rule-based systems*
    - *Evolutionary algorithms*
    - *Machine learning*

3.1 Computational Statistics

Computational Statistics models refers to computationally intensive statistical methods including Resampling methods (e.g., Bootstrap and Cross-Validation), Monte Carlo methods, Kernel Density estimation and other Semi and Non-Parametric Statistical methods, and Generalized Additive Models (Efron and Hastie, 2016; Wood, 2017). Examples include:

a) **Resampling methods** - a variety of methods for doing one of the following: i) estimating the precision of sample statistics using subsets of data (e.g. jack-knifing) or drawn randomly from a set of data points (e.g. bootstrapping); ii) exchanging labels on data points when performing significance tests (e.g. permutation tests); iii) validating models by using random subsets (e.g. repeated cross-validation);

b) **Monte Carlo methods** - a broad class of computational algorithms that rely on repeated random sampling to approximate integrals, particularly used to compute expected values (e.g. options payoff) including those meant for inference and estimation (e.g., Bayesian estimation, simulated method of moments);

c) **Kernel Density estimation** - are a set of methods used to approximate multivariate density functions from a set of datapoints; it is largely applied to generate smooth functions, reduce outlier effects and improve joint density estimations, sampling, and to derive non-linear fits;

d) **Generalized Additive Models** – a large class of nonlinear models widely used for inference and predictive modelling (e.g. time series forecasting, curve-fitting, etc.);
e) **Regularisation Methods** – Regularisation methods are increasingly used as an alternative to traditional hypothesis testing and criteria-based methods, for allowing better quality forecasts with a large number of features.

3.2 Complex Systems

A complex system is any system featuring a large number of interacting components (e.g. agents, processes, etc.) whose aggregate activity is nonlinear (not derivable from the summations of the activity of individual components) and typically exhibit hierarchical self-organization under selective pressures (Taylor, 2014; Barabási, 2016). Examples include:

a) **Cellular automata** - a collection of cells arranged in a grid, such that each cell changes state as a function of time according to a defined set of rules that includes the states of neighbouring cells;
b) **Agent-based models** - a class of computational models for simulating the actions and interactions of autonomous agents (individual or collective entities such as organizations or groups) with a view to assessing their effects on the system as a whole;
c) **Network-based models** - a complex network is a graph (network) with non-trivial topological features - features that do not occur in simple networks such as lattices or random graphs but often occur in graphs modelling of real systems; and
d) **Multi-Agent systems** – this subarea focus on formulating cooperative-competitive policies to a multitude of agents with the aim to achieve a given goal; this topic has significant overlap with Reinforcement Learning and Agent-based models.

3.3 AI and Machine Learning

There are broadly two classes of AI algorithms, which might be termed: *static algorithms* – traditional programs that perform a fixed sequence of actions; and *dynamic algorithms* – that embody machine learning and evolve. It is these latter 'intelligent' algorithms that present complex technical challenges for testing and verification, which will impact and demand further regulation.

This AI continuum of epistemological models spans three main communities:

a) **Knowledge-based** or heuristic algorithms (e.g. rule-based) - where knowledge is explicitly represented as ontologies or IF-THEN rules rather than implicitly via code (Giarratano and Riley, 1998);
b) **Evolutionary** or metaheuristics algorithms - a family of algorithms for global optimization inspired by biological evolution, using population-based trial and error problem solvers with a metaheuristic or stochastic optimization character (e.g. Genetic Algorithms, Genetic Programming, etc.) (Poli et al., 2008; Brownlee, 2011); and
c) **Machine Learning** algorithms - a type of AI program with the ability to learn without explicit programming, and can change when exposed to new data; mainly comprising *Supervised* (e.g. Support Vector Machines, Random Forest, etc.), *Unsupervised* (e.g. K-Means, Independent Component Analysis, etc.), and *Reinforcement Learning* (e.g. Q-Learning, Temporal Differences, Gradient Policy Search, etc.) (Hastie et al., 2009; Sutton and Barto, 2018). Russell and Norvig (2016) provide an in-depth view of different aspects of AI.

ML firstly subdivides into:

- **Supervised learning**: Given a set of inputs/independent variables/predictors $\mathbf{x}$ and outputs/dependent variables/targets $\mathbf{y}$, the goal is to learn a function $f(\mathbf{x})$ that approximates $\mathbf{y}$. This is accomplished by supervising $f(\mathbf{x})$, that is, providing it with examples ($\mathbf{x}_1$, $\mathbf{y}_1$), …, ($\mathbf{x}_\mathrm{n}$, $\mathbf{y}_n$) and feedback whenever it makes mistakes or accurate predictions.

- **Unsupervised learning:** Given several objects/samples $\mathbf{x}_1, \ldots, \mathbf{x}_\mathrm{n}$, the goal is to learn a hidden map $h(\mathbf{x})$ that can uncover a hidden structure in the data. This hidden map can be used to 'compress' $\mathbf{x}$ (aka dimensionality reduction) or to assign to every $\mathbf{x}_\mathrm{i}$ a group c_k (aka clustering or topic modelling).

- **Reinforcement learning**: Given an environment formed by several states $\boldsymbol{s}_1, \boldsymbol{s}_2, \dots, \boldsymbol{s}_n$, an agent, and a reward function, the goal is to learn a policy π that will guide an agent actions $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_k$ through the state space so as to maximize occasional rewards.

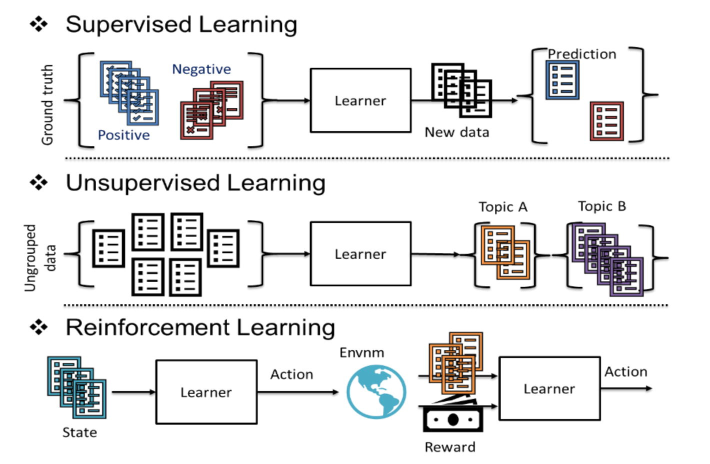


Figure 2: Main learning paradigms of Machine Learning.

Figure 2 provides an illustration of these key learning paradigms. Suppose a database of financial reports is available. If some of them have been historically labelled as positive and negative, we can leverage this to automatically tag future documents. This can be accomplished by training a Learner in a **Supervised** fashion. If these documents were unstructured, and spotting relations or topics is the goal (political events, economic data, etc.), a Learner trained in an **Unsupervised** manner can help uncover these hidden structures. Also, these documents can characterise the current state of the capital markets. Using that, a Learner can decide which actions should be taken in order to maximize profits, hedge against certain risks, etc. By interacting and gaining feedback from the environment, the Learner can Reinforce some behaviours so to avoid future losses or inaccurate decisions.

In addition to that, Deep Learning, Adversarial Learning, Transfer and Meta Learning are advanced new techniques enhancing Supervised, Unsupervised and Reinforcement learning. They are not only powering new solutions and applications (e.g. driverless vehicles, smart speakers, etc.) but they are making the resolution of previous problems cheaper, faster and more scalable. They tend also to be opaquer, making the issue of auditing and assurance more challenging. The second subdivision is:

- **Deep Learning** - deep learning algorithms attempt to model high-level abstractions in data by using multiple processing layers, with complex structures or otherwise, composed of multiple non-linear transformations. Hence, the mapping function we are attempting to learn can be broken down into several compositional operations $f(\mathbf{x}) = f_1 \circ f_2 \circ f_3 \circ \dots \circ f_n(\mathbf{x})$. Various deep learning architectures such as deep neural networks, convolutional deep neural networks, deep belief networks and recurrent neural networks have been applied to fields like computer vision, automatic speech recognition, natural language processing, audio recognition and bioinformatics where they have been shown to produce state-of-the-art results on various tasks (Goodfellow, et al., 2016; Chollet, 2017).

- **Adversarial Learning** - adversarial machine learning is a technique employed in the field of machine learning which attempts to 'fool' models through malicious input. More formally, assume a given input $\mathbf{x}$ associated to a label $\mathbf{c}$ and a machine learning model f such that $f(\mathbf{x}) = \mathbf{c}$, that is, f can perfectly classify $\mathbf{x}$. We consider $\mathbf{x}^*$ an adversarial example if $\mathbf{x}^*$ is indistinguishable from $\mathbf{x}$ and $f(\mathbf{x}) \neq \mathbf{c}$. Since they are automatically crafted, these adversarial examples tend to be misclassified more often than is true of examples which are perturbed by noise (Szegedy, 2013; Kurakin et al., 2016). Adversarial examples can be introduced during the training of models, making them more robust to attacks from adversarial agents. Typical applications involve increasing robustness in neural networks, spam filtering, information security applications, etc. (Huang et al., 2011).

- **Transfer/Meta Learning** – these two learning paradigms are tightly connected, as their main goal is to encapsulate knowledge learned across many tasks and transfer it to new, unseen ones. Knowledge transfer can help speed up training and prevent overfitting and can therefore improve the obtainable final performance. In Transfer Learning, knowledge is transferred from a trained model (or a set thereof) to a new model by encouraging the new model to have similar parameters. The trained model(s) from which knowledge is transferred is not trained with this transfer in mind, and hence the task it was trained on must be very general for it to encode useful knowledge with respect to other tasks. In Meta Learning the learning method (learning rule, initialization, architecture etc.) is abstracted and shared across tasks, and meta-learned explicitly with transfer in mind, such that the learning method generalises to an unseen task. Concretely, often in Transfer learning a pre-trained model is moved to a new task (Devlin et al., 2018; Radford et al., 2019), whilst in Meta learning a pre-trained optimizer is transferred across problems (Andrychowicz et al., 2016; Finn et al., 2017; Flennerhag et al., 2018). In both cases, the usual approach is to learn a Deep Neural Network that can be reused later, usually by stripping some of its terminal layers and creating an encoder-decoder to match the input and output for a task.

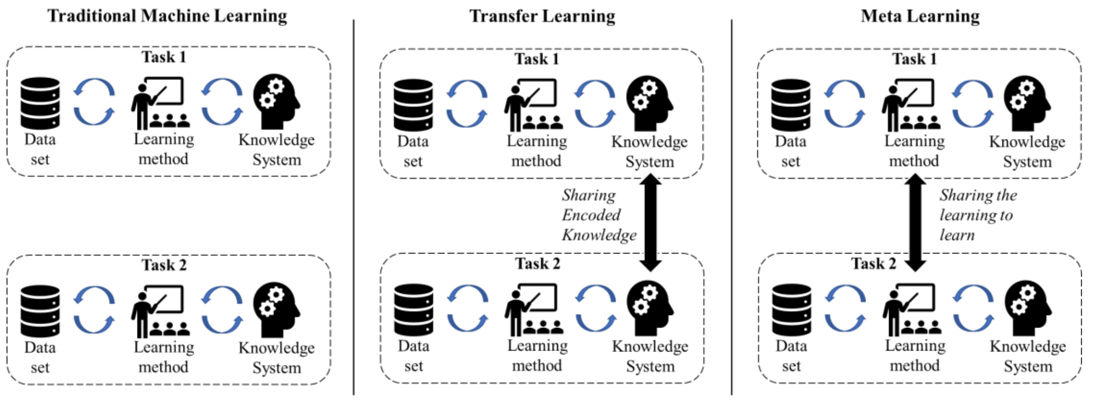

Traditional Machine Learning
Transfer Learning
Meta Learning
Task 1
Task 2
Data set
Learning method
Knowledge System
Sharing Encoded Knowledge
Sharing the learning to learn

Figure 3: Traditional ML versus Transfer Learning versus Meta Learning.

4. Main Verticals of Algorithm Auditing

In Computer Science, there is a growing engineering expertise overlapping with the Digital Ethics space (Floridi, 2018). Issues of explainability, fairness, privacy, governance, and robustness are now popular research themes among AI researchers – an area that is coming under the umbrella of “Trustworthy AI” (Brundage et al., 2020). From an engineering point of view, we believe that the most mature and impactful criteria are:

- **Performance and Robustness**: systems should be safe and secure, not vulnerable to tampering or compromising of the data they are trained on.
- **Bias and Discrimination**: systems should avoid unfair treatment of individuals or groups.
- **Interpretability and Explainability**: systems should provide decisions or suggestions that can be understood by their users, developers and regulators.
- **Algorithm Privacy**: systems should be trained following data minimization principles as well as adopt privacy-enhancing techniques to mitigate personal or critical data leakage.

The next subsection will deal with each one of these criteria.

4.1 Performance and Robustness

Performance and Robustness as a technical concept is closely linked to the principle of prevention of harm (EU-HLEG, 2019). Systems should neither cause nor exacerbate harm or otherwise adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity. Preventing harm can also entail consideration of the natural environment and of the living world. Most of the legal basis is established by an interaction between Regulatory Agencies, Professional Associations and Industry Trade Groups, where standards, rules and code of conducts are created:

- Financial algorithms: SEC, FCA, FSB, BBA, BIS
- Power systems: FERC, IEEE
- Electrical appliances: NIST, Nat Fire Protection Association, State Legislation

- Automotive sector: National Transportation Safety Board, Soc Auto Engineers
- And many others.

Algorithm Performance and Robustness is characterized by how effectively an algorithm can be deemed as safe and secure, not vulnerable to tampering or compromising of the data they are trained on. We can rate an algorithm's performance and robustness using four key criteria (EU-HLEG, 2019):

- **Resilience to attack and security**: AI systems, like all software systems, should be protected against vulnerabilities that can allow them to be exploited by adversaries, such as data poisoning, model leakage or the infrastructure, both software and hardware. This concept is linked with the mathematical concept of ***Adversarial Robustness*** (Carlini et al., 2019), that is, how would the algorithm have performed in the worst-case scenario? (e.g. how the algorithm would react during the 2008 Financial Crisis?). Mathematically, this can be expressed as:

  Adversarial risk[1]**:** $\mathbb{E}_{(\boldsymbol{x},\, \boldsymbol{y})\sim \boldsymbol{p}}\left[\max_{\delta\in\Delta(x)} L\big(y; f(x+\delta)\big)\right] \approx \underset{(x,y)\in D^{val}}{\text{mean}}\left[\max_{\delta\in\Delta(x)} L\big(y; f(x+\delta)\big)\right]$

- **Fallback plan and general safety**: AI systems should have safeguards that enable a fallback plan in case of problems. Also, the level of safety measures required depends on the magnitude of the risk posed by an AI system. This notion is strongly associated with the technical concept of ***Formal Verification*** (Qin et al., 2019), which in broad terms means: does the algorithm attends the problem specifications and constraints? (e.g. respect physical laws). One way to express this mathematically is:

  Verification bound[1]: $\mathbb{P}\big(F\big(x; f(x)\big) \leq 0\big) \approx \frac{\#\left(F\left(x^{nom}; f(x)\right)\leq 0\right)}{|S_{in}(x^{nom},\, \delta)|}$

- **Accuracy**: pertains to an AI system's ability to make correct judgements, for example to correctly classify information into the proper categories, or its ability to make correct predictions, recommendations, or decisions based on data or models. Accuracy as a general concept can be quantified by estimating the ***Expected Generalization Performance*** (Arlot and Celisse, 2010), which means that in general the question of *how well the algorithm works?* Is asked (e.g. in 7 out of 10 cases, the algorithm makes the right decision). Typically, the Expected Generalization Performance can be expressed by the following formula:

  Expected Loss[1]: $\mathbb{E}_{(x,\, y)\sim p}\big[L\big(y; f(x)\big)\big] \approx \underset{(x,y)\in D^{val}}{\text{mean}}\big[L\big(y; f(x)\big)\big]$

- **Reliability and Reproducibility**: a reliable AI system is one that works properly with a range of inputs and in a range of situations, whilst reproducibility describes whether an AI experiment exhibits the same behaviour when repeated under the same conditions. This idea is tied with the software engineering concept of ***Continuous Integration*** (Meyer, 2014), that is, is the algorithm auditable? (e.g. reliably reproduce its decisions).

4.2 Bias and Discrimination

Fairness as an ideal has been present in different manifestos and charters throughout history, gradually amplifying its outreach across the population, being the most recent and overarching in the UN *Universal Declaration of Human Rights* (1948). Most of the legal basis was developed after multiple public demonstrations, civil rights movements, etc. and are in many situations set or upheld at Constitutional levels. We can mention a few across different countries: US: Civil Rights Act (1957 and 1964), Americans with Disability Act (1990); UK: Equal Pay Act (1970), Sex Discrimination Act (1975),

[1] L: loss function; $\mathbb{E}$: expectation operator; y: output variable; x: input variable; $f(x)$: algorithm prediction/decision; p: sampling distribution of (x, y); D^{val}: holdout set of (x, y); $\Delta(x)$: set of feasible perturbations (δ) of x; F: specification mapping x and $f(x)$ in a real number, if $F\big(x; f(x)\big) \leq 0$ then, we say it is satisfied; $S_{in}(x^{nom},\, \delta)$: the set of all input x that are at most δ distant from x^{nom} ($S_{in}(x^{nom},\, \delta) = \{x\colon \left|\left|x - x^{nom}\right|\right|_{\infty} \leq \delta\}$); $\mathbb{P}$: probability measure.

Race Relations Act (1976), Disability Discrimination Act (1995), Equality Act (2010); and those enshrined in the constitutions of France, German, Brazil, and a many other countries. Indeed, it is suffice to say that notions of fairness appeal to substantive value claims rooted in differing philosophical approaches and traditions – as such there is often ambiguous interpretations of the word 'fairness'.

In AI and ML there are multiple sources of bias that explain how an automated decision-making process becomes unfair (EU-HLEG, 2019):

- **Tainted examples:** any ML system keeps the bias existing in the old data caused by human and societal biases (e.g. recruitment).
- **Skewed sample:** future observations confirm predictions made, which create a perverse, or self-justifying feedback loop (e.g. police record).
- **Limited features:** features may be less informative or reliably collected for minority group(s).
- **Sample size disparity:** training data coming from the minority group is much less than those coming from the majority group.
- **Proxies:** even if protected attributes are not used for training a system, there can always be other proxies of the protected attribute (e.g. neighbourhood).

To diagnose and mitigate bias in decision-making, we first need to differentiate between Individual and Group level fairness: (i) **Individual**: seeks for similar individuals to be treated similarly; and (ii) **Group**: split a population into groups defined by protected attributes and seeks for some measure to be equal across groups. There are multiple ways to translate these concepts mathematically (Chouldechova, 2016; Kleiberg et al., 2016; Corbett-Davies et al., 2017) and deciding which definition to use must be done in accordance with governance structures and on a case-by-case basis. Also, within Group fairness, it is possible to distinguish between the aim of Equality of Opportunity and Outcome. For example, using SAT score as a feature for predicting success in college:

- **Equality of Opportunity** worldview says that the score correlates well with future success and there is a way to use the score to correctly compare the abilities of applicants. A mathematical definition that is often used is the Average Odds Difference (Bellamy et al., 2018):

$$AOD = \frac{1}{2}\left[\left(FPR_{group\ A} - FPR_{group\ B}\right) + \left(TPR_{group\ A} - TPR_{group\ B}\right)\right]$$

with FPR and TPR representing the false and true positive rates, respectively. The underscore group A and group B reflect the conditioning of FPR and TPR to a given subset of the population analysed (e.g. group A could represent young individuals, and group B adult individuals).

- **Equality of Outcome** worldview says that the SAT score may contain structural biases so its distribution being different across groups should not be mistaken for a difference in distribution in ability. Statistical Parity Difference (Bellamy et al., 2018) is generally the most adopted form to represent this idea symbolically:

$$SPD = \frac{\mathbb{P}(\hat{y} = 1\ |group\ A)}{\mathbb{P}(\hat{y} = 1\ |group\ B)} \approx \frac{Freq(\hat{y} = 1\ |group\ A)}{Freq(\hat{y} = 1\ |group\ B)}$$

with $Freq$ representing the empirical frequency of positive/yes/etc. predictions $\hat{y}$ made by the model.

we can also list variations of both, like equal reliability (UK-CDEI, 2021). Calibration is also capable of perpetuating pre-existing biases. It should be noticed that fairness could be interpreted radically different in different environments and different countries and hence one deployment of a given algorithm may encounter several different fairness measurement barriers. Finally, it's perhaps worth noting somewhere that it's not mathematically possible to construct an algorithm that simultaneously satisfies all reasonable definitions of a "fair" or "unbiased" algorithm (Chouldechova, 2017).

4.3 Interpretability and Explainability

Being able to provide clear and meaningful explanations is crucial for building and maintaining users' trust in automated decision-making systems (Longo et al., 2020). This means that processes need to be transparent, the capabilities and purpose of systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested (EU-HLEG, 2019). The ultimate user benefits from being able to contest decisions, seek redress, and learn through user-system interaction; the developer also benefits from a transparent system by being able to "debug" it, uncover unfair decisions and from knowledge discovery.

Hence, the capabilities and purpose of algorithms should be openly communicated decisions explainable to those directly and indirectly affected timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher). In the US, credit scoring has a well-established right to explanation legislation via The Equal Credit Opportunity Act (1974). Credit agencies and data analysis firms such as FICO comply with this regulation by providing a list of reasons (generally at most four, per interpretation of regulations). From an AI standpoint, there are new regulations that gives the system's user the right to know why a certain automated decision was taken in a certain form -- Right to an Explanation – EU General Data Protection Regulation (2016).

In the context of AI and ML, Explainability and Interpretability are often used interchangeably. *Algorithm Interpretability* is about the extent to which a cause and effect can be observed within a system, and the extent an observer is able to predict what will happen, for a given set of input or algorithm parameters. *Algorithm Explainability* is the extent to which the internal mechanics of a ML (deep learning) system is explainable in human terms. In simple terms, Interpretability is about understanding the algorithm mechanics (without necessarily knowing why); Explainability is being able to explain what is happening in the algorithm.

There are multiple forms to generate and provide explanations based on an algorithmic decision-making system. Figure 4 presents the types and levels of Explainability: model-specific and agnostic, global and local (Hall, 2019; Molnar, 2019). Below we unwrap these concepts, as well as outline some technical solutions:

Model-specific (intrinsic): With model specific explainability, a model is designed and developed in such a way that it is fully transparent and explainable by design. In other words, an additional explainability technique is not required to be overlaid on the model in order to be able to fully explain its workings and outputs.

Model-agnostic (post-facto): With model-agnostic explainability, a mathematical technique is applied to the outputs of any algorithm including very complex and opaque models, in order to provide an interpretation of the decision drivers for those models.

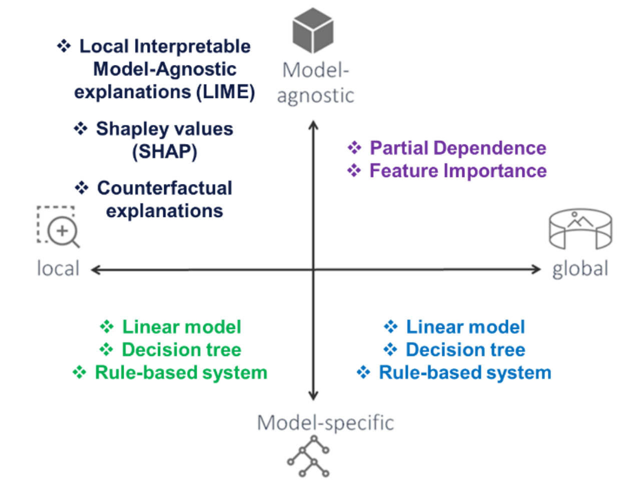


Figure 4: Types and levels of Algorithm Explainability.

Global: this facet focuses on understanding the algorithm's behaviour at a high/dataset/populational level. The typical user are researchers and designer of algorithms, since they tend to be more interested with the general insights and knowledge discovery that the model produces, rather than specific individual cases.

Local: this facet focuses on understanding the algorithm's behaviour at a low/subset/individual level. The typical user of local explanations are individuals being targeted by an algorithm, as well as members of the judiciary and regulators trying to make a case about potential discrimination.

It is important to note that the explainability requirements may be different for different regions and different use cases. This means that the same approach may not be applicable in all contexts of deployment of a given algorithm.

4.4 Algorithm Privacy

From the principles level, privacy is closely linked to the principle of prevention of harm (EU-HLEG, 2019): systems can cause or exacerbate adverse impacts due to asymmetries of power or information, such as between employers and employees, businesses and consumers or governments and citizens. Preventing harm demands bespoke data governance that covers the quality and integrity of the data used, its relevance considering the domain in which the algorithm will be deployed, its access protocols and the capability to process data in a manner that protects privacy. It is possible to group these issues in two key areas:

- **Privacy and data protection**: systems must guarantee privacy and data protection throughout a system's entire lifecycle (EU, 2016; Hind et al., 2018). This includes the information initially provided by the user and the one generated about the user over the course of their interaction with the system. Finally, protocols governing data access should be put in place, outlining who can access data and under which circumstances (Butterworth, 2018).
- **Model inferences**: the security of any system is measured with respect to the adversarial goals and capabilities that it is designed to defend against. In this sense, one needs to provide information about: (i) the level of access the attacker might have ('black-box' or 'white-box'); (ii) where the attack might take place (inference or training); and (iii) passive versus active attacks (De Cristofaro, 2020).

Therefore, the risk assessment of Algorithm Privacy can be disentangled in 'data', 'algorithm', and the interaction between both components. Below we outline the key methods available to assess risks coming from each of these elements:

- **Data**: the standard procedure to assess risks in this vertical is the Data Protection Impact Assessment (Bieker et al., 2016). This procedure has been legally formalized in many jurisdictions, such as in the EU, UK, Canada, California, Brazil, etc. In the UK, as shown in Figure 14, a qualitative rating can be provided depending on the perceived level of data protection. Another vector is data poisoning (Tan and Shokri, 2019), where an attacker maliciously manipulates the training data in order to affect the algorithm behavior.
- **Algorithm**: the key attack vector in this component is inferring model parameters and build 'knock-offs' version of it. To assess vulnerability, the auditor could apply techniques that aim to extract an (near-)equivalent copy or steal some functionalities of an algorithm (Ateniese et al., 2015; Tramèr et al., 2016; Orekondy et al., 2019).
- **Data-Algorithm interaction**: the attack vectors in this component are inferring about members of the population or about members of the training dataset through interactions with the algorithm. Attacks such as statistical disclosure (Dwork and Naor, 2010), model inversion (Fredrikson et al., 2015), inferring class representatives (Hitaj et al., 2017), membership and property inference (Shokri et al., 2017; Ganju et al., 2018; Melis et al., 2019) are different criteria that can be applied to an algorithm to assess levels of vulnerability.

4.5 Interactions and Trade-off Analysis

A vital area of exploration is the trade-off analysis between these verticals. As depicted by Figure 5, these verticals are not independent of each other – they overlap and interact. For example, with debiasing procedures affecting the model performance, global and local interpretation and, potentially, data minimization aspects. Having a clear understanding of what will be traded in consequence of improvements in one vertical is becoming less of a technological concern, and gradually more a requirement across a wide array of guidelines (EU-HLEG, 2019; ICO 2020; ICO-Turing, 2020). Above all, it presents the growing evidence that in the emerging area of Trustworthy AI hardly there is a solution only trade-offs to be managed. Though the practicalities of trade-off analysis demand context, nonetheless some general explorations, roadmaps and

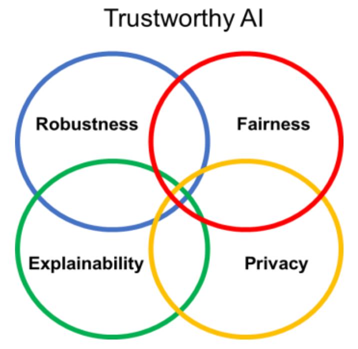


Figure 5: The overlaps between Algorithm Robustness, Fairness, Explainability and Privacy.

guidelines can be still be issued and performed. In what follows we explore some of these.

Explainability vs Robustness (accuracy): One that has been extensively explored by different authors and organizations (Koshiyama et al., 2020; ICO-Turing, 2020) is the Interpretability vs Accuracy trade-off – sometimes also presented as Explainability vs Performance trade-off. Figure 6 shows a typical depiction that can be found in many documents and papers. Prima facie, that is, looking only at the model function forms and training, the depiction is broadly accurate. However, such depiction is highly debatable in the light of data science practice, since it could be that a Linear model is the most accurate model, but due to massive pre-processing performed (e.g. nonlinear features, etc.) the explainability level has been drastically reduced.

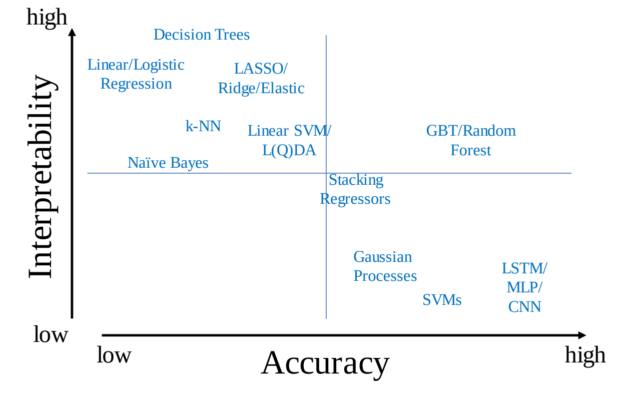


Figure 6: Algorithm selection trade-offs: model-specific Interpretability vs Accuracy.

Fairness vs Robustness: another trade-off well-explored in the literature is the Fairness (in the form of algorithm bias) and Robustness (in the form of algorithm performance) (Feldman et al., 2015; Kleinberg et al., 2016; Zafar et al., 2019). Figure 7 explores a typical chart about this trade-off. Every dot represents an algorithm setup (parameters, hyperparameters, etc.); the work of an algorithm designer is to identify the acceptable *boundaries of statistical bias and performance*, for example by adopting metrics like Statistical Parity and Accuracy. These boundaries can be identified by liaising with business and end users, and by analysing best practices, standards or regulations commonly adopted in the field of application. In the example depicted in Figure 7, the boundaries are set for -0.1 and 0.1 for Bias (Statistical Parity), and the Minimum Acceptable Performance of 0.53. From that, we can draw the region of algorithm configurations (or even models) that dwells within such limits. In this case, only 3 configurations are feasible from a Fairness vs Robustness point-of-view.

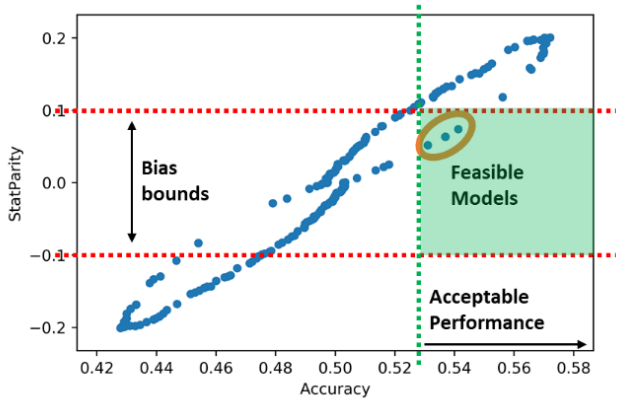


Figure 7: Algorithm selection trade-offs: Bias (Statistical Parity) vs Performance (Accuracy).

Explainability vs Privacy: prima facie the easier it is to interpret a model, the harder it is to conceal information or its judgement. Hence, in first sight, interpretability and privacy are negatively related. However, being able to explain a model's internal workings such as via Feature Importance charts can aid with Data Minimization (Goldsteen et al., 2020), a key pillar of Algorithm Privacy. Using Figure 8 as an example, if we set a threshold of 0.025 to the Feature Importance metric, we can reduce the number of variables being used from 20 to only 8. Knocking-off variables ease the explanation of a model judgements and will also reinforce to the end-users that their information is used in an efficient manner.

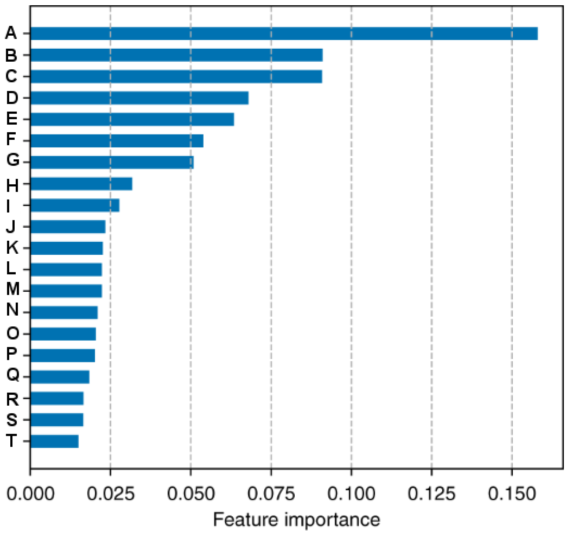


Figure 8: Algorithm selection trade-offs: Explainability (Feature Importance) vs Privacy (Data minimization).

Fairness vs Explainability: improving explainability of a system as a means to achieve greater transparency of its use acts as a positive driver to uncover inherent bias and discrimination to all its users and designers – se for instance Sharma et al. (2019). Figures 9a and 9b present examples using feature importance charts to understand the key drivers for a mortgage application processing algorithm. Figure 9a demonstrates that when we break down the feature importance chart per declared sex, we discover disparities in how the algorithm is making its judgement – even though we have not included this information as an input to the model. Loan amount and, particularly applicant income are significantly more relevant variables for female applicants than for male. We can perform a similar analysis, such as in Figure 9b, where a permutation importance method was used on Disparate Impact metric (male-female) to construct the feature importance chart. We uncover that there are disparities, as perceived in Figure 9a, particularly with the loan purpose, type and applicant income.

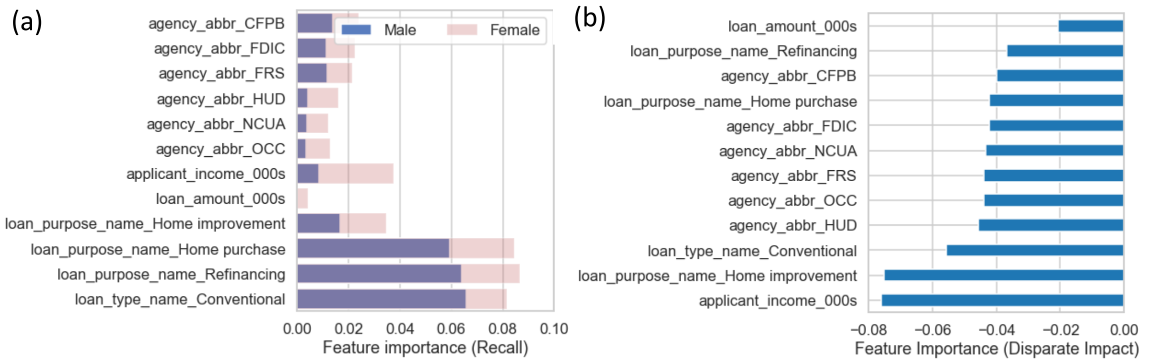


Figure 9: (a) Feature Importance chart with the breakdown per male and female groups. (b) Explainability (Feature Importance) based on a Bias (Disparate Impact) metric.

Interaction between all verticals: there are a few charts that can be crafted to display components of each vertical. Figure 10 displays one of such, where the key goal is to identify relevant variables and undertake data minimization. Relevant variables are defined as having a high impact in an algorithm Performance (Accuracy) and low impact in an algorithm Bias (Average Odds Difference) – both can be estimated by permutation importance using each as the loss metric. The variables J and K are key variables, meeting both criteria; the variables G and H could be eliminated since they do not affect much the model performance. Having this global understanding of an algorithm behaviour will become an unprecedented component to build and enhance the trustworthiness of an algorithm.

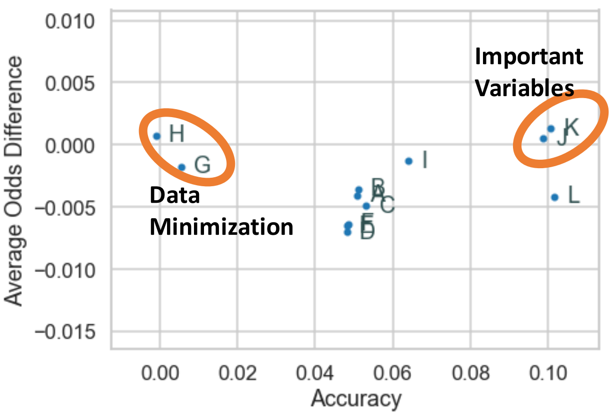


Figure 10: Interaction between all verticals. The values displayed were estimated by permutation importance using Accuracy and Average Odds Difference as loss metrics.

Two other interactions are worth briefly mentioning:

- **Robustness vs Privacy**: both criteria are strongly connected, with techniques coming from Privacy literature like Adversarial Testing (De Cristofaro, 2020) percolating to Robustness, and defence mechanisms built by the Robustness (Müller et al., 2019) community looping back.
- **Privacy vs Fairness**: respect for privacy and for fairness within the same system introduces the question of trade-offs between the two values. From the perspective of privacy, particularly in cases of personal data, the further a system is to anonymity the more ‘private’ it can be said to be. Conversely, in the case of fairness the concern is that systems perform equally for all protected attributes and, as such, systems need to be as transparent as possible for fairness to be assured. The tension between privacy and fairness becomes apparent, where a greater degree of privacy is likely to come at the price of fairness concerns (Kazim and Koshiyama, 2020a).

Notwithstanding the critical nature of trade-off analysis it should be noted that the intersection of all these areas is often impossible to achieve and not always desirable. Trade-offs should be seen as a way of finding an operational profile which is consistent with the needs of the application, rather than some abstract goal which needs to be achieved for a notion of “completeness”.

Future Investigations

One of the key challenges is to define what risks should be prioritized and measured. This is solved case-by-case, however a roadmap or toolkit could be developed to provide business users and developers with the right recommendation and areas to focus. In this perspective, future investigations could look at, given a specific algorithm, how to:

- Define the appropriate vertical or risks that should be prioritized as well as the right control levels for them:
    - *Bias and Discrimination*, such as when the algorithm will affect individuals or groups.
    - *Performance and Robustness*, such as when the algorithm can cause financial and reputational damage by not being statistically accurate or brittle.
    - *Interpretability and Explainability*, such as when the lack of understanding of the decisions being made, suggestions being provided, or recourse is needed.
    - *Privacy*, such as when the possibility of leakage of intellectual property or private information is a feasible event.
- Monitor metrics and recommend interventions depending on the phase, information provided, and the type of the project involved.
    - Development/procurement phase: provide recommendations of useful tools and techniques to include so that risks can be mitigated and avoided.
    - Deployment phase: request information about performance, bias and other metrics that is needed to assure that the risks are under control.

5. Levels of Access for Auditing

There are different levels of access that an auditor has during its investigation of an algorithm. In the scientific literature and technical reports, the commonplace is to categorize the knowledge about the system in two extremes: 'White-box' and 'Black-box'. In fact, the spectrum about the knowledge of a system is more of 'shades of grey' i.e. a continuum than this simple dichotomy. This additional nuance allows a richer exploration of the technologies available for assessment and mitigation, as well as the right level of disclosure that a certain business feels comfortable to engage.

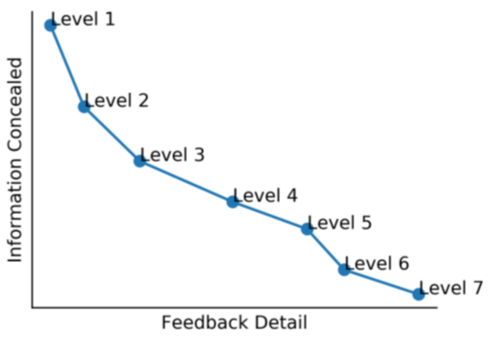


Figure 11: Information concealed versus feedback detail trade-off curve.

Hence, we can identify seven levels of access that an auditor can have of a system (Table 3). It ranges from 'Process-access' where only indirect observation of a system can be made to 'White-box' where all the details encompassing the model are disclosed. The levels in between are set by limiting the access to the components behind the learning process (e.g. knowledge of the objective function, model architecture, input data, etc.).

This categorization has the following two monotonic properties:

- **Detail**: accuracy and richness increase with levels;
- **Concealment**: information concealed decrease with levels;

In what follows we explore the trade-off: detail and concealment (Figure 11). It is worth mentioning that Level 7 access allows all the analysis of above levels, simply because we have full access to the algorithm. Conversely, analysis and techniques requiring Level 7 cannot be used at Level 6 without proper assumptions. Hence, Level 7 contains all the assessment, monitoring, and mitigation strategies of upper levels, with the report getting less detailed and inaccurate as levels increase.

Table 3. Landscape of Algorithm Auditing.

Dimension	*Level 1* *Process access*	*Level 2* *Model access* $f(.)$	*Level 3* *Input access* $f(x)$	*Level 4* *Outcome access* $f(x), y$	*Level 5* *Parameter control* $f_\theta(x), y$	*Level 6* *Learning goal* $L(f_\theta(x), y)$	*Level 7* *'White-box'*
Explainability	Checklist	Feature relevance Partial Dependency	Surrogate explanations	Accuracy of explanations	Stability of explanations	Model complexity	Documents and specific explanations
Robustness	Checklist	Adversarial attacks	Synthetic data	Concept drift analysis	Stability analysis	Stress-testing	Model selection and validation
Fairness	Checklist	Adversarial fairness	Bias in outcome	Bias in opportunity	Stability of bias metrics	Trade-off of bias and loss metric	Model selection and development
Privacy	Checklist	Statistical disclosure	Property and membership inference	Inversion attacks	Functionality stealing	Model extraction	Model security evaluation
Information Concealed	Very High	High	High	High/Medium	Medium	Medium	Low
Feedback Detail	Low	Medium	Medium	High/Medium	High	High	Very High
Typical Application	Sales Forecasting	Cyber-security	Recruitment	Credit-scoring	Facial Recognition	Algorithmic Trading	Self-driving vehicle
Appropriate Oversight	Guidelines	External Auditing/ Certification	External Auditing	External Auditing	External Auditing	Internal/ External Auditing	Internal Auditing

5.1 Level 7: 'White-box' Auditing

In the 'White-box' setup, the auditor knows all the details encompassing the model: architecture or type f, learning procedure and task objectives L, parameters θ, output y and input x data used to train and validate the model, and the access to perform predictions $f(.)$.

This level of access, very much identical to the system developer and business user have, allows the auditor to provide an accurate and richer feedback. Accurate because the whole assessment was performed using the actual system and based in little no assumptions; richer because the number of tests and recommendations that can be made range from the actual model selection, training, bias mitigation, validation, and security. It would be easier to assess mitigation strategies and provide actual information that can be more easily documented by the developers.

This level of access is more appropriate for internal auditors or in-house consultants, since this would demand an additional level of disclosure that may require non-disclosure, intellectual property sharing, data sharing, etc. agreements in place.

5.2 Level 6: Learning goal

In the Learning goal setup, the auditor knows most of the details encompassing the creation and purpose of the predictive system: learning procedure and task objectives L, parameters θ, output y and input x data used to train and validate the model, and the access to perform predictions $f(.)$.

From a modelling point of view, the auditor knows how to refit/re-learn the model using the actual incentives/objective function that it was trained on $L(f_\theta(x), y)$, but without knowing the model f is family (e.g. kernel method) or components (e.g. number of neurons).

This level of access allows the auditor to investigate an almost accurate picture of the system, without necessarily infringing much of the intellectual property. The feedback has a high degree of detail, with information of the model complexity, stress-testing, and trade-off analysis of bias, privacy, and loss being able to be performed without little to no assumptions. This level of access is enough to perform automated internal and external auditing, since the human involvement after setting up the APIs and environments, are considerably low.

5.3 Level 5: Parameter Manipulation

In the Parameter manipulation setup, the auditor can recalibrate/reparametrized the model, but has no information of its type or family, and what is the incentives/objective function it was built on. Hence, the auditor has access to parameters θ, output y and input x data used to train and validate the model, and the access to perform predictions $f(.)$.

This level allows explicitly the auditor to perform stability and perturbation analysis on the model f_θ. Hence, it enables to provide a reasonable feedback, covering particularly areas of how stable the system is performing, its judgements and the explanations being provided. Also, it would allow the auditor to assess the risk of functionality stealing from a privacy point of view. This level of access is relatively straightforward to implement via an API and can be easily automated for external auditing. The level of information known about the model nature is relatively low, allowing a low infringement of intellectual property or disclosures of other nature.

Also, since the auditor can reparametrize the model, and based on certain assumptions, the auditor can in practice retrain the model. It means that some analysis that would only be possible having Level 2 access could be performed in Level 3, but these assessments will only be considerate as 'assumption-based scenario-analysis' rather than actual Level 2 feedback.

5.4 Level 4: Outcome Access ('Gray-box')

In the Outcome access level, the auditor has the capacity to make predictive calls with the model using the actual input data, and to compare with outcome/output/target information. Therefore, the auditor has access to output y and input x data used to train and validate the model, and the access to perform predictions $f(.)$.

This setup is deemed by some authors as 'black-box', since the auditor does not know the parameters and architecture of the model. From a modelling perspective, a host of techniques are available to assess and operate at this level, most of them under the umbrella of 'model-agnostic' procedures (e.g. cross-validation, Shapley Values, etc.).

Since there are higher levels of non-access, we deem this level as 'Grey-box' since some information is still known to the auditor. With the available access, and based on a few assumptions, the auditor can perform concept drift analysis, investigate the accuracy of explanations, perform inversion attacks, and check bias from an equality of opportunity point of view (e.g. Equal Odds Difference). The auditor can also build baseline or competitor models to f.

Depending on the specifics, this yields a high to medium level of detail in the final feedback provided. From this level onwards, apart from data sharing agreements, there's little to no need to share intellectual property or development details. The level of automation that can be achieved and implemented make it possible to perform most analysis quicker and possibly in real-time.

5.5 Level 3: Input Data Access

In the Input data access setup, the auditor has the capacity to make predictive calls with the model using the actual inputs that has been used to train and validate it but cannot compare the predictions with the actual outcome data. That is, the auditor has only access to input x data used to train and validate the model, and the access to perform predictions $f(.)$.

The absence of outcome information y makes the problem of assessing the generalization behaviour of a model hard, particularly to assess its performance. Since only the predictions $f(x)$ are available, some analysis can still be performed, like computing bias from an equality of outcome perspective (e.g. Disparate Impact), property and membership inference, or creating surrogate explanations. Synthetic data, near to the actual distribution of the input x can be generated, allowing for an investigation of the model brittleness to gradual changes in the distribution.

5.6 Level 2: Model Access ('Black-box')

In the Model access level, the auditor has the possibility to make predictive calls with the model, but without having any information about the actual distributions of the input data. Some metadata could be shared, for example, the name of the variables, types, ranges, etc. Therefore, the auditor has only access to perform calls in $f(.)$ using some artificial input x^*.

This level of access entails the least amount of information disclosed to the auditor, since no data sharing agreements are needed. The level of automation that can be achieved is very high, since only API access is needed to perform the analysis. Most of the quantitative analysis performed is centered around an adversary setup, resembling the work of threat models performed in the privacy space. Adversarial attacks, adversarial evaluation of bias and discrimination (fairness), extracting feature relevance and partial dependency explanations, and different forms of privacy attacks (under the umbrella of statistical disclosure) are typical analysis that could be performed.

5.7 Level 1: Process Access

In the Process access setup, the auditor has no direct access to the algorithm, with its investigations and interventions occurring at the model development process. With the impossibility to perform calls at the model f, the auditor depends on checklists that can be partially qualitative and quantitative information. General and sector-specific guidelines issued by regulators and other governmental bodies supplemented by a combination of company/application-specific could form the body of the assessment. Probably for low-stakes and low-risk applications, this level of disclosure and feedback detail might be the most appropriate.

We believe that the above level of access scheme can be utilised by regulators and standard bodies in the context of balancing proprietary respect and risk, where context and sector sensitivities will be critical in deciding the level of access required.

Future Investigations

One of the key challenges is to specify which types of processes would be in play at each of these levels. For example, for each level how much interaction the auditor would need with the company being audited. One can imagine that for the deepest level of auditing, it may be necessary to first interview the key people in the company to ascertain their desires and goals for the operational parameters of the algorithms. Conversely, for the lowest level of auditing, simple checklists and self-assessment forms, may be sufficient. Perhaps also an automated tooling running over data and algorithms to produce high-level analysis.

On a more methodological side, it is difficult for a layman to assess which is the right level of auditing/oversight needed for a given algorithm. A roadmap or toolkit could be employed to set the right level of oversight needed for the AI application being developed or acquired:

- ‘Checklist-level’: when the risks are low, and no oversight is needed.
- ‘Black-box level’: when the risks are low-medium and little oversight is needed.
- ‘Grey-box level’: when the risks are medium, and some oversight is needed.
- ‘Glass-box level’: when the risks are medium-high and full oversight is necessary.

Even though it is application specific, undoubtedly the sector and company size will affect the level of auditability available. We conjecture that startups will be more inclined to be open when they are being procured by large organizations. We are not sure about middle size or small organizations though. Questions related to protection of intellectual property, business sensitive information, etc. will be raised and may demand a clear protocol from a legal point of view. Finally, the need for auditing to be reproducible and scalable are crucially important.

6. Mitigation Strategies

Mitigation strategies are a set of techniques employed to address issues highlighted in the assessment part of algorithm auditing. They consist of specific procedures that can be used in conjunction in order to enhance an algorithm performance or solve issues like algorithm debiasing or establishing surrogate explanations. To some extent they act as ‘add-ons’ to certain stages of model development, and hence, demand a retraining and reassessment of the model -- Figure 12 establishes this feedback loop. We can highlight two types of mitigation procedures:

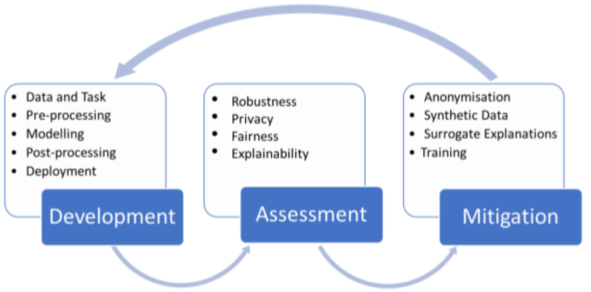


Figure 12: Feedback loop: from model development, assessment and mitigation, to redevelopment, reassessment and re-mitigation.

- **Human**: all procedures that involve how algorithm developers design, collaborate, reflect and develop algorithms. These procedures can involve (re)training, impact assessment, etc.
- **Algorithm**: all methodologies that can be applied to improve an algorithm current outcome.

These approaches are not in conflict and one solution may end-up using both procedures in concert. In this section we explore mainly the mitigation strategies that can be employed to improve an algorithm Robustness, Explainability, Privacy and Fairness.

Performance and Robustness: each technical criterion listed in section 4.1. embodies several technical mitigations strategies (Table 4). These technical strategies can aid the analyst in measuring the expected generalization performance, detecting concept drifts, avoiding adversarial attacks, and having best practices in terms of systems development and algorithm deployment.

Table 4. Mapping technical criteria and solutions for algorithm robustness and performance.

Criteria	Technical Solution
Expected generalization performance	▪ **Cross-validation** (Arlot and Celisse 2010): k-fold-cv, leave-one-out, etc. ▪ **Covariance-penalty** (Efron and Hastie, 2016): Mallow's C_p, Stein Unbiased Risk Estimator ▪ **Concept drift** (Lu et al, 2018; Escovedo et al., 2018): gradual mitigation, abrupt correction, pre-emptive detection
Adversarial robustness	▪ **Evasion attacks**: fast gradient sign method (Huang et al., 2017), DeepFool (Moosavi-Dezfooli, 2016), etc. ▪ **Defence**: label smoothing (Müller et al., 2019), variance minimization (Guo et al., 2017), Thermometer Encoding (Buckman et al., 2018), etc.
Formal verification	▪ **Complete**: Satisfiability Modulo Theory (Bunel et al., 2018; Barrett et al., 2018), Mixed Integer Programming (Tjeng and Tedrake, 2017) etc. ▪ **Incomplete**: Propagating bounds (Huang et al., 2019), Lagrangian Relaxation (Dvijotham et al., 2018), etc.
Reliability and reproducibility	▪ **Code versioning**: Git (Github), Mercurial (BitBucket), etc. ▪ **Reproducible analysis**: Binder, Docker, etc. ▪ **Automated testing**: Travis CI, Scrutinizer CI, etc.

Explainability and Interpretability: most interpretability and explainability enhancing strategies concentrate at in processing and post-processing stage (Table 5). We can split the procedures mainly in the model-specific and model-agnostic axis, with all model-specific approaches being able to provide global and local explanations by-design (in-processing). Model-agnostic procedures act as a post-hoc 'wrapper' around an algorithm, with some techniques only focusing on local explanations (e.g. LIME) or global explanations (e.g. Partial Dependency plots). The mitigation strategies need to take into account the use case domain and level of risk, the organisation's risk appetite, all applicable regulation and laws, and values/ethical considerations.

Table 5. Modelling stage and different technical solutions for algorithm explainability and interpretability.

Stage/Method	Technical Solution
In-processing/ Model-specific	▪ **Rule-based explanations:** decision trees, rule-induction methods ▪ **Model's coefficients**: linear regression, linear discriminant analysis ▪ **Nearest prototype:** k-nearest-neighbour, Naïve-Bayes
Post-processing/ Model-agnostic	▪ **Surrogate explanations:** LIME (Ribeiro et al., 2016), Explainable Boosting Machines (Nori et al., 2019), PIRL (Puiutta et al., 2020) ▪ **Perturbation:** Gradient-based Attribution Methods (Ancona et al., 2017)**,** Permutation Importance (Breiman, 2001), SHAP (Lundberg and Lee, 2017) ▪ **Simulation analysis (what-if?):** counterfactual explanations and algorithmic recourse (Wachter, 2017; Karimi et al., 2020)

Bias and Discrimination: regardless of the measure used, algorithm bias can be mitigated at different points in a modelling pipeline: Pre-processing, In-processing, and Post-processing (Bellamy et al., 2018). Table 6 presents a snapshot of different methodologies to mitigate bias in AI systems.

Table 6. Modelling stage and different technical solutions for algorithm bias and discrimination.

Stage	Technical Solution
Pre-processing	▪ **Reweighing subjects** (Kamiran et al., 2012) ▪ **Oversampling minority group** (Iosifidis and Ntoutsi, 2018) ▪ **Disparate Impact Remover** (Feldman et al., 2015) ▪ **Learning Fair Representations** (Zemel et al., 2013)
In-processing	▪ **Adversarial Debiasing** (Zhang et al., 2018) ▪ **Fairness constraint** (Zafar et al., 2019; Donini et al., 2018) ▪ **Counterfactual Fairness** (Kusner et al., 2017)
Post-processing	▪ **Calibrated equality of odds** (Pleiss et al., 2017) ▪ **Reject option classification** (Kamiran et al., 2012)

Algorithm Privacy: from an engineering standpoint, there are emerging privacy-enhancing techniques to mitigate personal or critical data leakage. These techniques can act in different moments of the system development: (i) during pre-processing stage by feature selection, dataset pseudo-anonymisation and perturbation; (ii) during in-processing by using federated learning, differential privacy, and model inversion mitigation; and (iii) deployment by implement rate-limiting and user's queries management. Table 7 presents these methods and key references.

Table 7. Modelling pipeline and different technical solutions for algorithm privacy.

Stage	Technical Solution
Pre-processing	▪ **Data Minimisation by Dim Reduction** (Goldsteen et al., 2020) ▪ **Dataset (Pseudo)-Anonymisation** (Neubauer and Heurix, 2011) ▪ **Dataset Perturbation** (Kargupta et al., 2005)
In-processing	▪ **Federated Learning** (McMahan and Ramage, 2017; Kim et al., 2019) ▪ **Differential Privacy** (Abadi et al., 2016; Dwork et al., 2015) ▪ **Model Inversion Mitigation** (Fredrikson et al., 2015) ▪ **Data Poisoning Defence** (Steinhardt et al., 2017)
Deployment	▪ **Rate-limiting** ▪ **User's queries management**

Future Investigations

On the mitigation point generally, one assumes that the auditor would recommend the mitigation procedures that would need to be applied in order to address identified issues. Perhaps they would recommend a range of options. Perhaps they would require a given mitigation even. Different levels would demand different timelines and activities. Figure 12 fleshes out the general perspective, but one could explore in more detail what could be done in different levels, such as:

- Level 7 – 'White-box' level
    - starts with an interview for goals and context with the development and business team;
    - deep dive to examine the system with the development team;
    - write a report with the details of the system and the business problem it is aiming to solve as well as recommendations to improve it;
    - mitigation strategies are implemented, and the system is re-developed;
    - another audit is performed to assure that the key performance metrics are attained.
- Level 1 – 'Checklist' level

- starts with a self-assessment performed by the team developing the system;
- depending on the stage of development and verticals to be prioritized, recommendations of interventions or metrics are reported;
- a final documentation is issued with possible monitoring and checkpoints for further assessment.

7. Assurance Processes

The broader outcome of an auditing process is to improve confidence or ensure trust of the underlying system. After assessing the system, and implementing mitigation strategies, the auditing process assesses whether the system conforms to regulatory, governance and ethical standards. Providing assurance, therefore, needs to be understood through different dimensions and steps needs to be taken so that the algorithm can be shown to be trustworthy. Below we list the key points that embodies the assurance process:

- **General and sector-specific assurance:** broad national regulation and standards (provided agents such as NIST, UK-ICO, EU, etc.) with sector specific ones, such as in financial services (e.g. SEC, FCA, etc.), health (e.g. NIH, NHS, etc.), real estate (e.g. RICS, IVS, USPAP), etc.
- **Governance:** from two aspects, namely technical assessments (robustness, privacy, etc.) and impact (risk, compliance, etc.) assessments.
- **Unknown Risks:** discussing risk schemes and highlighting 'red teaming', which is used to mitigate unknown risks.
- **Monitoring Interfaces:** outlining risk assessments and the use of 'traffic-light' user friendly monitoring interfaces.
- **Certification:** the numerous ways in which certification may occur, such as certification of a system or AI engineers, etc.
- **Insurance**: a subsequent service to emerge as a result of assurance maturing.

Figure 13 outlines the steps towards assurance: combining governance and impact assessments with audit and technical assessment; finding equivalent standards and regulations in the sector/end-application; generating a document/audit trail that will feed into a certification and insurance as part of assurance. We expand each point in the forthcoming subsections.

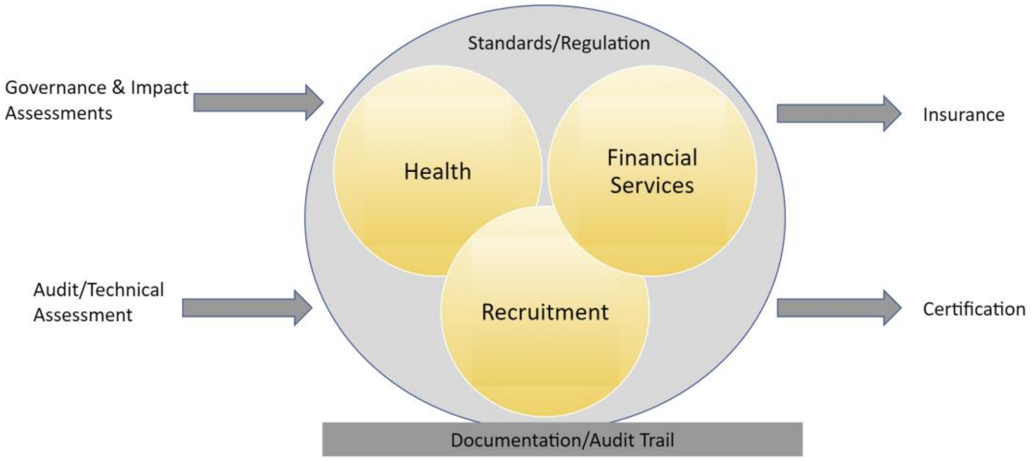

Standards/Regulation
Governance & Impact Assessments
Health
Financial Services
Insurance
Recruitment
Audit/Technical Assessment
Certification
Documentation/Audit Trail

Figure 13. Diagram outlining the steps towards assurance: combining governance and impact assessments with audit and technical assessment; finding equivalent standards and regulations in the sector/end-application; generating a document/audit trail that will feed into a certification and insurance as part of assurance.

7.1 General and sector-specific

The satisfaction of a particular standard - ex. certification, auditability, etc. - will become mandatory. We read this from the growing calls for AI, ML and associated algorithms to be responsibly developed and appropriately governed (European Commission, 2020; UK Committee on Standards in Public Life; UK-ICO, 2020; ICO, 2020). We anticipate that standards will be both general and sector specific:

- **General Standards:** the guidance (which may or may not be legally codified) will encompass broad dimensions such as privacy, explainability, safety and fairness, and these will be set by institutions and bodies with non-sector specific remits (e.g. the UK's Information Commissioner's Office). Developments in this space are becoming more concrete. For instance, in the recent 'Explaining decisions made with AI' , the Information Commissioner's Office & The Alan Turing Institute (ICO-Turing, 2020) advises on how organisations can explain the processes, services and decisions delivered or assisted by AI to those that are affected by such decisions - the guidance outlines explanations in terms of who is responsible, data choices and management, fairness considerations, safety and impact.
- **Sector Standards:** sector specific guidance already exists. These address idiosyncrasies of application. For example, the UK's Financial Conduct Authority is leading in the debate about standards of AI systems in financial services (Mueller and Ostmann, 2020), UK's Care Quality Commission in ML development for medical diagnostic services (UK-CQC, 2020), US's Department of Defence in the defence space (US-DOD, 2020). In addition to sector specific regulators issuing guidance, sectors themselves are developing their own standards and approaches to best practice. Recruitment is an example of this (UK-CDEI, 2020). Application-specific, like the US's NIST for Facial Recognition (NIST, 2020), is a promising avenue.

7.2 Governance

Governance can be divided into two broad streams, namely technical and non-technical:

- **Non-technical governance:** concerns systems and processes that focus on allocating decision makers, providing appropriate training and education, keeping the human-in-the-loop, and conducting social and environmental impact assessments. The issue of accountability and sector specific particularities dominate the current debate; here what is being referred to is:
    - who will be liable if something goes wrong (processor, controller, user) i.e. the allocation of responsibility;
    - what current legislation like GDPR, financial regulations, etc. have to say on a case-by-case basis; and
    - differences between countries and economic blocks.

  Within this context there is also the Algorithmic Impact Assessment literature, which calls for doing a Data Protection Impact assessment when algorithms are used (Reisman et al., 2019; Koshiyama and Engin, 2019; Canada, 2019; Kaminski and Malgieri, 2020). Additionally, there are calls for AI impact assessments that address issues of human-rights, social and environmental concerns (McGregor et al, 2019; EU-HLEG, 2018).

- **Technical governance:** concerns systems and processes that render the activity of the technology itself accountable and transparent. This touches upon ethical-by-design and technical auditing (involving creation of quantitative metrics for tracing and tracking decisions, making the technologies accessible for verification and accountability). The main dimensions of technical auditing that will be surveyed are (European Commission, 2020):
    - **Robustness and Performance**: systems should be safe and secure, not vulnerable to tampering or compromising - including the data they are trained on. Key concepts in this dimension are resilience to attack and security, fallback plan and general safety, accuracy/performance, and reliability and reproducibility.
    - **Bias and Discrimination** (Fairness): systems should use training data and models that account for bias in data, to avoid unfair treatment of certain groups. By bias we mean, for example, yielding more false positives to a group in relation to another (young people vs older people, etc.). Key sources of bias include tainted or skewed examples, limited features, sample size disparity and proxies to protected attributes.

- **Explainability and Interpretability**: systems should provide decisions or suggestions that can be understood by their users and developers. Key techniques in this space are individual/local explanations, population/global explanations, model-agnostic and model-specific interpretations.
- **Privacy**: systems should be trained following data minimization principles as well as adopt privacy-enhancing techniques to mitigate personal or critical data leakage. Key concepts in this area are data protection, quality, accuracy, integrity and access to data and decisions.

7.3 Monitoring interfaces

A risk-based approach, as observed in the European Commission's White Paper on Artificial Intelligence and the German Data Ethics Commission (German Data Ethics Commission, 2019), outlines two distinct notions of risks:

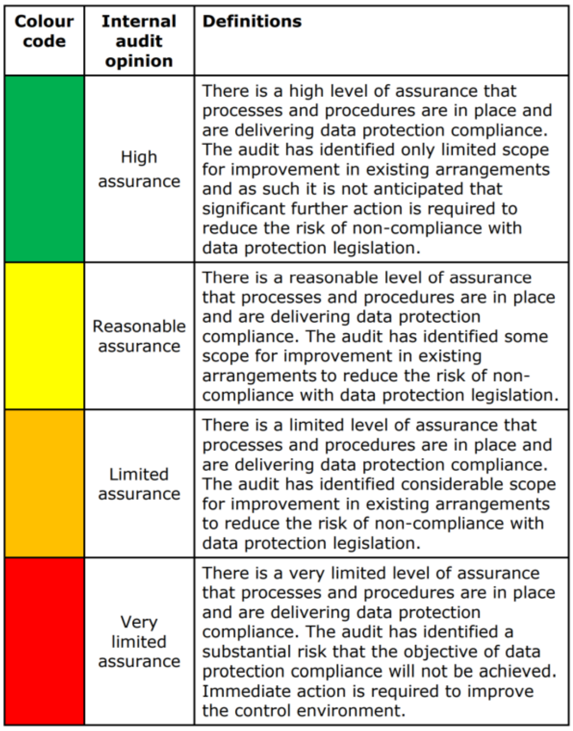
- **Sectorial**: where high-risk is identified with respect to things such as healthcare, transport, energy, and, parts of the public sector (ex. asylum, social security and employment services).

Table 8. Risk matrix outlining concerns and mitigation between technical/non-technical dimensions and known/unknown risks.

Audit and Impact Assessment	*Known*	*Unknown*
Technical	Bias/fairness; safety; explainable; accessibility; data protection; trails; verification; comprehensibility.	Breakdown/robustness; nature of hack (theft; DOS)
Non-Technical	Governance; oversight; whistle blowing; lack of education (education/training); authorisation.	Trust; reputational; psychological and social impact; loss of skills;

We note that all these sectors have the commonality of human impact i.e. whether a service, instruction, decision, etc. impacts on a human user and citizen. This is a broad, abstracted and blanketed approach, that is highly likely to result in two things, i. **risk aversion**, and ii. a**utonomous systems will be a high cost venture**. For example, a simple healthcare booking chatbot can become economically unfeasible because it falls under healthcare. Similarly, in the context of high-risk high-reward a risk-based approach based upon sector will **discourage potentially high-positive impact algorithmic systems** (ex. Medical applications of AI has significant risk and lifesaving potential). As such we believe this will stifle innovation (which is what the EU white paper calls for).

- **Use**: The second notion of risk introduced is that 'where use means that significant risk is likely to arise (risk of injury, death or significant material or immaterial damage)'.



Colour code	Internal audit opinion	Definitions
🟩	High assurance	There is a high level of assurance that processes and procedures are in place and are delivering data protection compliance. The audit has identified only limited scope for improvement in existing arrangements and as such it is not anticipated that significant further action is required to reduce the risk of non-compliance with data protection legislation.
🟨	Reasonable assurance	There is a reasonable level of assurance that processes and procedures are in place and are delivering data protection compliance. The audit has identified some scope for improvement in existing arrangements to reduce the risk of non-compliance with data protection legislation.
🟧	Limited assurance	There is a limited level of assurance that processes and procedures are in place and are delivering data protection compliance. The audit has identified considerable scope for improvement in existing arrangements to reduce the risk of non-compliance with data protection legislation.
🟥	Very limited assurance	There is a very limited level of assurance that processes and procedures are in place and are delivering data protection compliance. The audit has identified a substantial risk that the objective of data protection compliance will not be achieved. Immediate action is required to improve the control environment.

Figure 14. UK's Information Commissioner's Office has a colour coded 'Assurance Rating' for Data. Available at: https://ico.org.uk/for-organisations/audits/

A concern with this categorisation of risk is that it is unclear how unintended consequences can be assessed. We argue that risk can be thought of in terms of known and unknown risk, and technical and non-technical risk (presented in 'Risk Matrix' Table 8 (Kazim and Koshiyama, 2020b).

Given the problems referred to above and the vagueness of 'risk' in these calls, drawing from industry precedence, intuitive performance dashboard stop-light interfaces have been proposed – these will facilitate monitoring of performance over time (Bellotti et al 2020; Brundage et al 2020). With Green, Amber and Red representing high-performance, satisfactory-performance and poor-performance respectively. Furthermore, from a regulatory and standards standpoint, the UK's Information Commissioner's Office has a colour coded 'Assurance Rating' for Data (Figure 14). A stop light system can be used in several ways, like in the deployment phase where green, amber and red can be read in terms of how a system is performing in accordance with the purpose of its deployment. Within the context of assurance and audit the respective colours can be read in terms of high-performing/compliant (green), low-performing/compliant (amber), non-compliant (red).

7.4 Unknown risks

Foundational to safety is that steps should be taken and procedures in place that *prevent harm*. This preventative approach requires that risks are anticipated in order to ensure that the chances of them occurring are minimised and if they do occur then the impact is minimal. In order to so this, risk assessments are performed - in the context of the above, we can think of two kinds of risk assessment:

- **Technical audits:** conducted in the development phase and for live monitoring.
- **Impact assessments:** conducted before deployment and to design mitigation strategies.

Note that in Table 8 the known technical and non-technical risks are covered by audit and impact assessment; this leaves the unknown technical and non-technical risks. Within the literature one approach to addressing such unknown risks is through 'Red Teaming' (Brundage et al 2020) [24]:

- **Red Teaming**: a systematic attempt to probe, expose flaws and weaknesses in a system, process, organisation etc., both technical and non-technical, is undertaken. The 'red team exercise' assumes the persona of a hostile agent, with the hope that in exposing thereunto unanticipated weaknesses i.e. unknown risks, the risk mitigation can be improved.

Although there will still be unknown risks, through such activities it is hoped that best practice can be established; notwithstanding proprietary issues, this can be facilitated through knowledge-transfer (via publication of methods to probe 'attack' and mitigate) (Brundage et al 2020) [24].

7.5 Certification

Certification is the part of the assurance process by confirming that a system, process, organisation, etc. satisfies a particular standard. It is typically intertwined with regulatory requirements. However, certification can also be granted by industry bodies or other recognised authorities. We read certification as a final 'stamp' or confirmation, which can be achieved via providing evidence and proving that a system, process, organisation has satisfied the set standards. Certification may come in a number of forms, including:

- **Certification of a system:** here, likely to align with national regulatory and standard bodies, the use of AI i.e. the systems and governance, is may be certified as trustworthy or responsible. This may be akin to the granting of an organisational licence.
- **Sector specific certification:** here it is possible that sector standard bodies and regulators issue their own sector specific certification.
- **Certification of a responsible agent:** good practice and industry standards within the context of data protection has led to the position of a 'Data Protection Officer', and by analogy something akin to a 'Responsible AI Officer' may emerge. These officers may be certified.
- **Certification of Algorithm Engineers:** Here the AI engineers may be certified, as though being granted a licence or admission into an accreditation organization (c.f. Trade Association).

Another possibility is that certification may be issued for specific aspects of a system; here certifications for **Robustness**, **Explainability**, **Privacy** and **bias and discrimination** may be issued.

7.6 Insurance

Closely related to assurance is the insurance of algorithms. It is possible that this will become a significant risk mitigation requirement for companies engaged in automation, and as such a significant market for insurers. We envision that this will align closely with Explainability and Algorithm Auditing in accordance with regulation and standards. Pricing such contracts will demand understanding of the risks involved in each vertical of the algorithm system (Robustness, Bias, etc.) as well as indemnity insurance for high-risk sectors or high-risk end-application.

Future Investigations

Certification is a topic that demands a section of its own. Questions related to: should one certificate be issued for the whole process, or parts of the system? What could be shared with third parties to declare that the algorithms have been audited and verified. This brings us into the area of certificating authorities – who are they and what are their roles – how do they (if at all) differ from the auditor.

Accountability roles is a topic that also demands another section, separating the obligations of each of the players in the supply chain - the one that commissions the algorithm, the designer, the coder, the tester, the operator, and so on. One can use analogies such as a comparison with the General Product safety regulations, where the obligations are primarily on the manufacturer of goods, but the distributor and retailers have lesser but serious obligations to ensure safety.

8. Final Remarks

This work is a first step towards understanding the key components underlying Algorithm Auditing. We provide a list of definitions and a taxonomy, since this area is a combination of research done mostly in silos, such as bias and discrimination, robustness, explainability, and privacy. Our goal with this paper is to instigate the debate in this novel area of research and practice and to kick-start that debate with a robust set of areas, processes and strategies. Translating concepts such as Accountability, Fairness, Transparency, into engineering practice is non-trivial, with its impact perceived in design choices, algorithms to be used, delivery mechanisms and built infrastructure. This demands a full integration with respect to governance structures with real-time algorithm auditing.

We foresee that a new industry is emerging, Auditing and Assurance of Data and Algorithms, with the remit to professionalize and industrialize AI, ML and associated algorithms. Since the magnitude of the challenge will increase year-on-year for the foreseeable future, this industry will increasingly demand human capital (AI/Digital ethicists, data scientists), RegTech-inspired solutions and business models (Treleaven and Batrinca, 2017), and (thought-)leadership from concerned regulators, politicians, NGOs, and academics.

Below we highlight related questions (which have not been covered extensively in this document):

- **AI, ML and Algorithm Ethics**: with the proliferation of AI research and deployment, along with high-profile cases of harm, awareness of the social impact and ethical implications of AI has risen to the fore. What is now referred to as 'AI ethics' or 'trustworthy AI' or 'responsible AI' is the body of literature that has resulted because of this consciousness and debate (Hagendorff, 2020). The field of AI ethics has undergone three broad phases (Kazim and Koshiyama, 2020c): principles, ethical-by-design approach, and – indeed the current phase – as concerned with the need to standardise and operationalise the AI ethics discipline.
- **Legal Status of Algorithms**: there is the growing discussion regarding algorithms and the law, in particular concerns regarding fairness and automation (Wachter et al., 2020) in the Judiciary concerning the 'status of algorithms in Law'. In Law, as we know, companies have the rights and obligations of a person. Algorithms are rapidly emerging as artificial persons: a legal entity that is not a human being but for certain purposes is legally considered to be a natural person (Treleaven et al., 2019). The argument is that since algorithms are doing or intermediating business (agency) with humans, companies and even other algorithms they also need to have the status of an artificial person in Law.

Finally, to reiterate some points made previously, there is a growing demand for a tool that could assist procurement, information security and internal developers of AI applications to self-assess a solution and flag if:

- They are performing *low-risk applications* and should go ahead.
- They are performing *medium-risk applications* and should provide more information and implement mitigations strategies.
- They are performing *high-risk applications* and should go through a review process before deploying their solution across business.

We believe that this could be the interface that would connect the verticals and the different mitigation strategies.

Acknowledgements

Adriano Koshiyama and Philip Treleaven would like to acknowledge Cisco Research Centre for their research grant (#2019-207109, 3696). We would like to thank Lindsay Edwards, Pratik Ajmera, Harshwardhan Prasad, Nick Firoozye, Graca Carvalho, Lisa Whiting, Magdalena Lis, Mark Durkee, Alastair Pearson Roger Taylor, Umar Mohammed, and Michael Cheetham for their valuable support. We would also like to acknowledge the different stakeholders from the UK CQC, FCA and ICO for the multiple interactions, publications and activities that have informed this emerging area of research.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016, October). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308-318).

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2017). Towards better understanding of gradient-based attribution methods for deep neural networks. arXiv preprint arXiv:1711.06104.

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., ... & De Freitas, N. (2016). Learning to learn by gradient descent by gradient descent. In Advances in neural information processing systems (pp. 3981-3989).

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics surveys, 4, 40-79.

Ateniese, G., Mancini, L. V., Spognardi, A., Villani, A., Vitali, D., & Felici, G. (2015). Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. International J of Security and Networks, 10(3), 137-150.

Barabási, A. L. (2016). Network science. Cambridge university press.

Barrett, C., & Tinelli, C. (2018). Satisfiability modulo theories. In Handbook of Model Checking (pp. 305-343). Springer, Cham.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Nagar, S. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943.

Belloti, A., Hand, D. J., Khan, S. (2020). Predicting Through a Crisis. Second White Paper. Validate AI.

Bieker, F., Friedewald, M., Hansen, M., Obersteller, H., & Rost, M. (2016, September). A process for data protection impact assessment under the european general data protection regulation. In Annual Privacy Forum (pp. 21-37). Springer, Cham.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Brownlee, J. (2011). Clever algorithms: nature-inspired programming recipes. Jason Brownlee.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... & Maharaj, T. (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213.

Buckman, J., Roy, A., Raffel, C., & Goodfellow, I. (2018, February). Thermometer encoding: One hot way to resist adversarial examples. In International Conference on Learning Representations.

Bunel, R. R., Turkaslan, I., Torr, P., Kohli, P., & Mudigonda, P. K. (2018). A unified view of piecewise linear neural network verification. In Advances in Neural Information Processing Systems (pp. 4790-4799).

Butterworth, M. (2018). The ICO and artificial intelligence: The role of fairness in the GDPR framework. Computer Law & Security Review, 34(2), 257-268.

Canada (2020). Directive on automated decision-making. Ottawa (ON): Government of Canada (modified 2019-02-05.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., ... & Kurakin, A. (2019). On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705.

Chollet, F. (2017). Deep Learning with Python. Manning Publications co.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2), 153-163.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 797-806).

De Cristofaro, E. (2020). An Overview of Privacy in Machine Learning. arXiv preprint arXiv:2005.08679.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., & Pontil, M. (2018). Empirical risk minimization under fairness constraints. In Advances in Neural Information Processing Systems (pp. 2791-2801).

Dvijotham, K., Stanforth, R., Gowal, S., Mann, T. A., & Kohli, P. (2018, March). A Dual Approach to Scalable Verification of Deep Networks. In UAI (Vol. 1, p. 2).

Dwork, C., & Naor, M. (2010). On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. Journal of Privacy and Confidentiality, 2(1).

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). Generalization in adaptive data analysis and holdout reuse. In Advances in Neural Information Processing Systems (pp. 2350-2358).

Efron, B., & Hastie, T. (2016). Computer age statistical inference (Vol. 5). Cambridge University Press.

Escovedo, T., Koshiyama, A., da Cruz, A. A., & Vellasco, M. (2018). DetectA: abrupt concept drift detection in non-stationary environments. Applied Soft Computing, 62, 119-133.

EU (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)

EU-HLEG. (2019). Ethics guidelines for trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

European Commission (February 2020). White Paper on Artificial Intelligence: A European approach to excellence and trust. Available at: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 259-268).

Finn, C., Abbeel, P., & Levine, S. (2017, August). Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 (pp. 1126-1135). JMLR. org.

Flennerhag, S., Moreno, P. G., Lawrence, N. D., & Damianou, A. (2018). Transferring knowledge across learning processes. arXiv preprint arXiv:1812.01054.

Floridi, L. (2018). Soft ethics, the governance of the digital and the General Data Protection Regulation. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 20180081.

Fredrikson, M., Jha, S., & Ristenpart, T. (2015, October). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1322-1333).

Fredrikson, M., Jha, S., & Ristenpart, T. (2015, October). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1322-1333).

Ganju, K., Wang, Q., Yang, W., Gunter, C. A., & Borisov, N. (2018, January). Property inference attacks on fully connected neural networks using permutation invariant representations. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (pp. 619-633).

German Data Ethics Commission. (2019). Opinion of the Data Ethics Commission. Bmjv.de. Available at: www.bmjv.de/SharedDocs/Downloads/DE/Themen/ Fokusthemen/Gutachten_DEK_EN_lang.pdf (Accessed: 30 August 2020).

Giarratano, J. C., & Riley, G. (1998). Expert systems. PWS publishing co..

Goldsteen, A., Ezov, G., Shmelkin, R., Moffie, M., & Farkash, A. (2020). Data Minimization for GDPR Compliance in Machine Learning Models. arXiv preprint arXiv:2008.04113.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Guo, C., Rana, M., Cisse, M., & Van Der Maaten, L. (2017). Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 1-22.

Hall, P. (2019). An introduction to machine learning interpretability. O'Reilly Media, Incorporated.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Hind, M., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Olteanu, A., & Varshney, K. R. (2018). Increasing trust in AI services through supplier's declarations of conformity. arXiv preprint arXiv:1808.07261.

Hitaj, B., Ateniese, G., & Perez-Cruz, F. (2017, October). Deep models under the GAN: information leakage from collaborative deep learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 603-618).

https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/

Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I., & Tygar, J. D. (2011, October). Adversarial machine learning. In Proceedings of the 4th ACM workshop on Security and artificial intelligence (pp. 43-58). ACM.

Huang, P. S., Stanforth, R., Welbl, J., Dyer, C., Yogatama, D., Gowal, S., ... & Kohli, P. (2019). Achieving verified robustness to symbol substitutions via interval bound propagation. arXiv preprint arXiv:1909.01492.

Huang, S., Papernot, N., Goodfellow, I., Duan, Y., & Abbeel, P. (2017). Adversarial attacks on neural network policies. arXiv preprint arXiv:1702.02284.

ICO (2020). Guidance on AI auditing framework: Draft guidance for consultation. https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

ICO-Turing (2020). Explaining decisions made with AI Information. Information Commissioner's Office & The Alan Turing Institute.

Iosifidis, V., & Ntoutsi, E. (2018). Dealing with bias via data augmentation in supervised learning scenarios. Jo Bates Paul D. Clough Robert Jäschke, 24.

Kaminski, M. E., & Malgieri, G. (2020, January). Multi-layered explanations from algorithmic impact assessments in the GDPR. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 68-79).

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1-33.

Kargupta, H., Datta, S., Wang, Q., & Sivakumar, K. (2005). Random-data perturbation techniques and privacy-preserving data mining. Knowledge and Information Systems, 7(4), 387-414.

Karimi, A. H., Schölkopf, B., & Valera, I. (2020). Algorithmic Recourse: from Counterfactual Explanations to Interventions. arXiv preprint arXiv:2002.06278.

Kazim, E., & Koshiyama, A. (2020c). A High-Level Overview of AI Ethics. Available at SSRN.

Kazim, E., & Koshiyama, A. (2020a). The Interrelation Between Data and AI Ethics in the Context of Impact Assessments. Available at SSRN.

Kazim, E., and Koshiyama, A. Human Centric AI: A Comment on the IEEE's Ethically Aligned Design (April 13, 2020b). Available at SSRN: https://ssrn.com/abstract=3575140 or http://dx.doi.org/10.2139/ssrn.3575140.

Kim, H., Park, J., Bennis, M., & Kim, S. L. (2019). Blockchained on-device federated learning. IEEE Communications Letters.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807.

Koshiyama, A., & Engin, Z. (2019). Algorithmic Impact Assessment: Fairness, Robustness and Explainability in Automated Decision-Making.

Koshiyama, A., Firoozye, N., & Treleaven, P. (2020). Algorithms in Future Capital Markets. Available at SSRN 3527511.

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In Advances in neural information processing systems (pp. 4066-4076).

Longo, L., Goebel, R., Lecue, F., Kieseberg, P., & Holzinger, A. (2020, August). Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 1-16). Springer, Cham.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. IEEE Transactions on Knowledge and Data Engineering, 31(12), 2346-2363.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in neural information processing systems (pp. 4765-4774).

McGregor, L., Murray, D., and Ng, V. (2019). International human rights law as a framework for algorithmic accountability. International & Comparative Law Quarterly, 68(2), 309-343.

McMahan, B., & Ramage, D. (2017). Federated learning: Collaborative machine learning without centralized training data. Google Research Blog, 3.

Melis, L., Song, C., De Cristofaro, E., & Shmatikov, V. (2019, May). Exploiting unintended feature leakage in collaborative learning. In 2019 IEEE Symposium on Security and Privacy (SP) (pp. 691-706). IEEE.

Meyer, M. (2014). Continuous integration and its tools. IEEE software, 31(3), 14-16.

Molnar, C. (2019). Interpretable machine learning. Lulu. com.

Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2574-2582).

Mueller, H. and Ostmann, F. (2020). AI transparency in financial services – why, what, who and when?. Financial Conduct Authority. Available at: https://www.fca.org.uk/insight/ai-transparency-financial-services-why-what-who-and-when

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help?. In Advances in Neural Information Processing Systems (pp. 4694-4703).

Neubauer, T., & Heurix, J. (2011). A methodology for the pseudonymization of medical data. International journal of medical informatics, 80(3), 190-204.

NIST (2020). Ongoing Face Recognition Vendor Test (FRVT). National Institute of Standards and Technology. Available at: https://pages.nist.gov/frvt/reports/11/frvt_11_report.pdf

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223.

Orekondy, T., Schiele, B., & Fritz, M. (2019). Knockoff nets: Stealing functionality of black-box models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4954-4963).

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In Advances in Neural Information Processing Systems (pp. 5680-5689).

Poli, R., Langdon, W. B., McPhee, N. F., & Koza, J. R. (2008). A field guide to genetic programming. Lulu. com.

Puiutta, E., & Veith, E. (2020). Explainable Reinforcement Learning: A Survey. arXiv preprint arXiv:2005.06247.

Qin, C., O'Donoghue, B., Bunel, R., Stanforth, R., Gowal, S., Uesato, J., ... & Kohli, P. (2019). Verification of non-linear specifications for neural networks. arXiv preprint arXiv:1902.09592.

Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019). Better language models and their implications. OpenAI, 2018b. URL https://openai. com/blog/better-language-models.

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2019). Algorithmic impact assessment: a practical framework for public agency accountability. AI Now Institute.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " 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 (pp. 1135-1144).

Rushby, J. (1988). Quality measures and assurance for AI (Artificial Intelligence) software.

Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited.

Sharma, S., Henderson, J., & Ghosh, J. (2019). Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. arXiv preprint arXiv:1905.07857.

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 3-18). IEEE.

Steinhardt, J., Koh, P. W. W., & Liang, P. S. (2017). Certified defences for data poisoning attacks. In Advances in neural information processing systems (pp. 3517-3529).

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

Tan, T. J. L., & Shokri, R. (2019). Bypassing backdoor detection algorithms in deep learning. arXiv preprint arXiv:1905.13409.

Taylor, S. (Ed.). (2014). Agent-based modeling and simulation. Springer.

Tjeng, V., & Tedrake, R. (2017). Verifying neural networks with mixed integer programming. arXiv:1711.07356, 945-950.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction apis. In 25th USENIX Security Symposium (USENIX Security 16) (pp. 601-618).

Treleaven, P., & Batrinca, B. (2017). Algorithmic regulation: automating financial compliance monitoring and regulation using AI and blockchain. Journal of Financial Transformation, 45, 14-21.

Treleaven, P., Barnett, J., & Koshiyama, A. (2019). Algorithms: law and regulation. Computer, 52(2), 32-40.

UK Committee on Standards in Public Life (2020). Artificial Intelligence and Public Standards: report. Lord Evans of Weardale KCB DL. Available at: https://www.gov.uk/government/publications/artificial-intelligence-and-public-standards-report

UK-CDEI (2020). Review into bias in algorithmic decision-making. Centre for Data Ethics and Innovation. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/957259/Review_into_bias_in_algorithmic_decision-making.pdf

UK-CQC (2020). Using machine learning in diagnostic services. UK's Care Quality Commission. Avaialble at: https://www.cqc.org.uk/sites/default/files/20200324%20CQC%20sandbox%20report_machine%20learning%20in%20diagnostic%20services.pdf

US-DOD (2020). DOD Adopts Ethical Principles for Artificial Intelligence. Available at: https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech., 31, 841.

Wood, S. N. (2017). Generalized additive models: an introduction with R. Chapman and Hall/CRC.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019). Fairness Constraints: A Flexible Approach for Fair Classification. J. Mach. Learn. Res., 20(75), 1-42.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, February). Learning fair representations. In International Conference on Machine Learning (pp. 325-333).

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335-340).