AI and Ethics
https://doi.org/10.1007/s43681-023-00343-z

ORIGINAL RESEARCH

Algorithmic audits of algorithms, and the law

Erwan Le Merrer[1] · **Ronan Pons**[2] · **Gilles Tredan**[3]

Received: 24 April 2023 / Accepted: 4 September 2023

Abstract
Algorithmic decision-making is now widespread, ranging from health care allocation to more common actions such as recommendation or information ranking. The aim to audit these algorithms has grown alongside. In this article, we focus on external audits that are conducted by interacting with the user side of the target algorithm, and hence considered a black box. Yet, the legal framework in which these audits take place is mostly ambiguous to researchers developing them: on the one hand, the legal value of the audit outcome is uncertain; on the other hand, the auditors' rights and obligations are unclear. The contribution of this article is to articulate two canonical audit forms to law, to shed light on these aspects: 1) the first audit form (we coin the `Bobby` audit form) checks a predicate against the algorithm, while the second (`Sherlock`) is looser and opens up to multiple investigations. We find that: `Bobby` audits are more amenable to prosecution, yet are delicate as operating on real user data. This can lead to rejection by a court (notion of admissibility). `Sherlock` audits craft data for their operation, most notably to build surrogates of the audited algorithm. It is mostly used for acts for *whistleblowing*, as even if accepted as proof, the evidential value will be low in practice. 2) these two forms require the prior respect of a proper right to audit, granted by law or by the platform being audited; otherwise, the auditor will be also prone to prosecutions regardless of the audit outcome. This article thus highlights the relation of current audits with law, to structure the growing field of algorithm auditing.

Keywords Algorithmic decision-making · Machine learning · Audits · Law

1 Introduction

Law is supposed to hold service providers and their algorithms accountable. In particular, *decision-making algorithms*, , are now widespread [1]. They directly face users, and govern large portions of our lives (from apparently subtle decisions such as recommendations, to more life-changing ones such as criminal justice or health care allocation [2]). The legal perspective on algorithms, especially of online platforms, is indeed evolving into strong legal frameworks. For example, at a European level, to protect the fundamental rights of European residents,[1] or to frame artificial intelligence systems[2]; this explicitly shows the willingness to better regulate algorithms.

The IT perspective and the nascent field of audits Computer scientists and engineers are designing and developing algorithms that process information and that can have an important impact on society [4, 5]. For improving and refining these, developers operate a feedback loop on data fed as inputs to the algorithm, and on the corresponding algorithm results (output accuracy for instance).

✉ Erwan Le Merrer
erwan.le-merrer@inria.fr

[1] Univ Rennes, Inria, CNRS, Irisa, Rennes, France

[2] UT1 Capitole, Université d'Ottawa & ANITI, France & Canada, Capitole, France

[3] LAAS/CNRS, Toulouse, France

[1] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC COM/2020/825 final.

[2] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT).

Published online: 27 September 2023

Considering an exterior viewpoint (the viewpoint of users or regulators) that observes or audits the behavior of remote algorithms is less frequent. A so-called *black box* approach to algorithms can be dated back to Moore's tests of black box automata in 1956 [7]. Relatively recent and sporadic works instead placed this viewpoint at the service of algorithmic auditing, to allow users to gain some understanding on the algorithmic decisions they are facing [1, 8–18]. In particular, these nascent forms of algorithmic audits can also constitute a prerequisite to enabling platform regulation [19] : if a state wants to enforce some behavior, means for verification are mandatory.

And the law? There is a blind spot regarding the current development of algorithmic audit techniques: what are their legal groundings? little (if any) mention is made in research works of the legal consequences of the conducted audit. In a nutshell, two fundamental questions are unaddressed: (i) what are the legal risks taken by an auditor, and (ii) can the outcome of an audit be used against the platform in court? This lack is a structural problem, as the question to whom the discovered issues must be addressed is central. Possible recipients can then be the general public (*e.g.*, through the act of *whistleblowing*), or justice.

In both cases, legal aspects are at stake: the auditor is likely to have violated the terms of service of the audited algorithm, displayed on the platform. What are then the consequences she faces? Or the consequences of the findings in a trial? If the auditor is in her own right to perform certain actions, what are the audit steps that will prevent acceptance of scientific proof in the eyes of justice?

Contributions In an attempt to shed light on the relation between audits and the law, we first propose to bind two novel prototypes of audit algorithms (that encapsulate state-of-the-art audit algorithms) to specific law perspectives. Since the law is by definition a sovereign prerogative to each state in the world, we will take the French law system as an instance and example in our presentation. We will as well include European perspectives from the global law frameworks currently in progress. This yields two salient points: 1) the simplest form of audit (we coin the `Bobby`audit) is easily usable in court, yet it is more delicate as it leverages real data (which can be a cause of rejection if all care has not been taken regarding laws such as the *GDPR*: the General Data Protection Regulation[3] 2) The most complex audit form (we coin the `Sherlock`audit) is less problematic as it crafts data to input in the algorithm of the platform; yet, it is far more difficult to bring to justice due to possibly lower probative value, leaving it in priority for whistleblowing. Finally, we review the recent proposal for a "vetted researchers" status by the European Union, and its consequences for future audits.

2 Two canonical algorithms capturing algorithmic audit schemes

There is a growing diversity of audits in the recent literature [1, 8–18]. Each one spans a specific behavior of a specific platform with its own methodology. In an attempt to structure this novel field, we introduce a set of fundamental distinctions that allows to separate these audits into two broad categories that distinguish audits both on their technical approach, and on their relevance for a potential trial.

Let us take the parallel with police work, tasked to check the application of the law. On the one hand, `Bobby`-family audits are tasked to tour the audited algorithm evaluating a well-defined characteristic of the platform, similar to a policeman tasked to tour a district to fine car parking infringements. Key to this approach is the existence of a logical predicate that very precisely defines the desirable (resp. undesirable) behavior of the audited algorithm, similar to the set of driving regulations that precisely define what is a correctly parked car. On the other hand, `Sherlock`-family audits target a deeper and loosely defined characterization of some aspect of the audited algorithm, similar to an inspector trying to elucidate some crime. Such an approach typically requires some interpolation to provide a general analysis based on some observed examples of the algorithm behavior.

2.1 Context and terminology

Let us consider the following use case: an individual (or a group of) hereafter named *the auditor* seeks to study the behavior of some algorithm executed remotely by some platform hereafter named a *target platform* or the *target algorithm*. Auditors are completely external to the target platform, and hence can only interact with the public side of the algorithm as regular users would.

This context is tailored to represent the typical context of a black box audit, where auditors are simple users interested in understanding or evaluating the behavior of the platform they use. This context can also capture situations in which the competent authorities have no specific access to the algorithm and want to verify the compliance of this behavior with some regulations.

Concretely, we wish to capture a spectrum of use cases ranging from informal citizen-driven audits (see *e.g.*, COMPAS) to academic research work on auditing platforms. All those situations cover the same high-level steps: an auditor writes some code to (i) request the target platform (either through some API, or through its web interface directly),

[3] Règlement (UE) 2016/679.

(ii) parse and collect the target algorithm answers, and (iii) publicize some analyses based on the collected data.

Formally, let A be the target algorithm. Let X (resp. Y) be the input (resp. output) space of A. Like regular users, auditors can only submit some request $x \in X$ to A, and then record the corresponding result $A(x) \in Y$.

2.2 The Bobby audit form

This is the simplest category of audits: in this form, an infraction is constituted by an input set (that is data existing in a dataset), to which corresponds a (problematic) collected output.

Algorithm 1 The Bobby audit

Input: A an algorithm to audit. $A : X \mapsto Y$, L a propositional formula over an input dataset $X_L = \{x_1 \dots x_l\} \subset X$ and corresponding outputs $Y_l = \{A(x_1), \dots, A(x_l)\} \subset Y$
Output: *True* if the behavior is illegal, *False* otherwise

```
infraction = False
for X not exhausted do
    Pick x_1, ..., x_l
    Collect Y_1 = A(x_1), ..., Y_l = A(x_l)
    if not L(X_1, Y_1), ... or L(X_l, Y_l) then
        infraction = True ;                          // A does not verify L
        break
end
return infraction ;                                  // Boolean on violation of L
```

Propositional formula L In the pseudo-code presented on Algorithm 1, the central component is the definition of the propositional formula L to be checked against the audited algorithm. In its definition, L encodes the desirable property one wants to observe. More precisely, L is a propositional formula defined over a set of input/output couples of the target algorithm that constitute the variables of the proposition. In L, those variables are linked by logical operators, such that L is well formed and has a *truth value*: L is either true or false.

As an illustrating example, imagine A is the algorithm that is in use in an online flight search platform that allows users to seek and book flights. For each request (departure and destination locations belonging to the IATA list, and dates), it provides the user with an ordered list of flights f_1,f_2, …. Assume the platform operating A declares that it ranks the resulting flights according to their cost. Such an assertion can be easily converted to a propositional formula that can be evaluated over any couple of flights f_i,f_j: $L_{cost}(f_i,f_j) := i \leq j \Rightarrow cost(f_i) \leq cost(f_j)$. Such declaration can be audited with a Bobbyaudit that regularly requests A to verify if L holds. In our example, an input for which L is violated is a couple of two returned flights f_a,f_b, such that f_a is more expensive and yet presented before f_b (formally: $a < b \wedge cost(f_a) > cost(f_b)$). If such violating input is found, the algorithm stops and reports this observation.

2.2.1 Bobby approaches in the literature

We now illustrate some examples of Bobby from the literature.

Cookies and transparency consent form auditing The GDPR and ePrivacy Directive recently set that European users must explicitly consent to non-necessary data collection, in general stored as a consent cookie on the users' computers. The checking of this rule is easily auditable: Matte et al. [25] implemented a crawler that (i) visits a target website without any interaction and (ii) detects the writing of a cookie registering consent by the target website. In that typical Bobbyaudit, the input space X can be all the target's webpages, and the predicate is L ="not a positive consent cookie".

The detection of "fairwashed" explanations Online services are now increasingly proposing to explain the main factors driving some of their automated decisions. The rationale is for them to gain the trust of the general public. Nevertheless, there is a possibility that the provided explanations are faked (or coined fairwashed [9]) to justify a discriminative decision. In work by Le Merrer et Tredan [26], so-called incoherent pairs are looked for; these are two conflicting explanations that yet give the same decision, and are the sign of a fairwashed explanation by the audited algorithm. This can be written as $L = \{\nexists In = ((a, white), (a, black)) \in X^2 \, s.t. \, f(a, white) \neq f(a, black)\}$. This mimics associations that are performing discrimination tests at the entrance of clubs for instance.

Copyright auditing Some forms of audits are dedicated to identifying copyright infringements by remote algorithms (*i.e.*, algorithms being executed without permission). The audit result is a Boolean answer on whether or not the remote algorithm is indeed the one that is suspected. This relates to the field of *watermarking*, where an algorithm is queried, and returns specific outputs if it is indeed the one suspected of infringement [27]. Here, the inputs used as queries are specifically designed to operate as identification keys. The predicate resemble $L = \{\forall In \in K_g, f(In) == g(In)\}$, with K_g being the watermark.

Skin color or gender bias audits Multiple studies fit in this class: to take a precise example, Buolamwini et al. [10] benchmark three commercial gender classifier systems with an intersectional dataset (skin color/gender). Gender classification accuracies are compared: the paper notes that *e.g.*, classification is 8.1% to 20.6% worse on female than male subjects and 11.8% to 19.2% worse on darker than lighter subjects.

Interestingly, this paper first constructs a dataset of faces that have balanced gender and skin types. Assuming this dataset is standardly recognized as a good benchmark for face classification, one could imagine a `Bobby`approach that targets any face classification algorithm using as input D : the standardized dataset. To implement the predicate function from Algorithm 1, classification results obtained on D could be compared for instance against a 60% disparity ratio [28]. For any partition of D into a gender/skin type subset D_s and its complementary $D_{\bar{s}}$, one has to compute the target algorithm's accuracy: $a_s = 1/|D_s|. \sum_{i,label(i)\in D_s} A(i) == label(i)$. For each partition s covered by the dataset, one then evaluates the predicate $L_s = \frac{a_s}{a_{\bar{s}}} > 0.6$.

Diversity in search engine results

Urman et al. [29] track several search engines, to audit source diversity and search concentration. This is achieved by submitting a static list of 62 keywords. As for bias, the final predicate can take the form of a simple rule such as one where the diversity at search engine B must be at least 0.6 the one at B for instance.

2.2.2 Limits of `Bobby`

The `Bobby` forms of audits are bound to verify a predicate L over an input space X. Given an input budget N (i.e., the number of different input queries sent to A), three outcomes are possible. Among them, two are to be considered as potential limitations.

Either some input $c \in X$ violating L is found (i.e., $L(c)$ is false). In this case, an infraction has been found, and simply exhibiting c and its corresponding answers $A(c)$ is sufficient to establish the infraction to L committed by A. Either no input violating L was found, and two sub-cases are to be distinguished:

1) $N > |X|$: the whole input space of A is exhausted, and no violation has been found. It is then legitimate to conclude that L is respected by A. Unfortunately, current algorithms have input spaces that are either unbound (e.g., with inputs being floats) or have a size orders of magnitude larger than typical auditing budgets N (e.g., few hundred queries for an input size of $3 \times 224 \times 224$ corresponding to images [30]).

2) $N < |X|$: no violation of L has been found within the budget N. In this quite common case, the auditor is left with a statistical guarantee but no definite answer. While the precise nature of the statistical guarantee depends on the specifics of the study (e.g., how the input space is sampled using the available input dataset), such an assertion typically translates that the empirical probability $\hat{P}_{\overline{L(c)}}$ of finding an input c that violates L is less than $1/N$.

A second limitation of `Bobby`resides in the production of inputs to query the target algorithm. We here stress that inputs from X (l. 3) can belong to a dataset (e.g., an image dataset in [10]), or be formed by public data. Thus, this line hides a wide variety of situations that are both heterogeneous with respect to the technical difficulty of generating X, and heterogeneous with respect to legal consequences.

Regarding the technical difficulty first: target algorithms working on simple inputs, like a text request on a search engine, might be queried with datasets that are easily collected. On the other hand, targeting algorithms taking more complex inputs like a video, a resume or a medical record, is certainly a challenge for the auditor to constitute these datasets.

A similar legal heterogeneity also resides in constituting datasets. Consider a flight search engine whose input is awaited to be by airport names along with some future travel date: both are public data that the auditor can rightfully use. On the other hand, testing a job recommendation engine that would match candidate resumes with job offers might require the auditor to submit resumes found on the web for which privacy rights (among others) exist and requires specific conditions to be processed (i.e., consent). In this second case, a judge might consider that the auditor had no right to use these data, and hence refuse to consider the conclusions drawn.

A third limitation of `Bobby`is the reliance on a propositional formula L: while some desirable behaviors of target algorithms can easily be converted into propositional formulas (e.g., if declared age is below 9, do not show ads), some others are intrinsically impossible to convert to such a logical statement (e.g., if declared age is below 9 do not propose shocking videos). This limits the applicability of `Bobby` audits to some specific, well-defined and desired properties.

2.3 The Sherlock audit form

This second form of audit algorithms is more elaborate, flexible and does not focus on the verification of a single propositional formula. These audits target a different set of infractions that, instead of relying on the collected outputs alone, rather relate to the general behavior of the audited algorithm. To pursue the parallel with policemen: building the case for, say, a murder requires our policeman to come up with a complete narrative (including motives, absence of alibi, etc.).

A Sherlock audit also needs to collect interaction sessions that characterize the target behavior (i.e., sequences of input and output pairs), and based on these examples, to *interpolate* on the behavior of the target algorithm. A typical Sherlockaudit thus contains two phases: a first phase that builds a local model of the target algorithm (hereafter named a *surrogate*), and a second phase that analyses the surrogate to extract its desired properties.

The algorithm is presented in Algorithm 2. The input crafting operation (l. 3) is here central: it pertains to a general plan to extract specific information from A, to create an accurate surrogate for A on the auditor's machine. This local surrogate S is then analyzed locally. While Bobby audits simply evaluate a predicate L, the analysis of S is much more open-ended (ranging from identifying shocking corner cases to characterizing the internal logic and comparing against other equivalents of A). To capture this diversity in a compact way, we define the set of *Acceptable* situations the auditor would refer to when conducting such an analysis.

Algorithm 2 The Sherlock audit

Input: A: an algorithm to audit. $A : X \mapsto Y$, N: A budget (maximum number) of queries

```
1 I ← find input to A
  /* Build a surrogate S to A                                          */
2 while n < N do
3 |   craft a new I_n ∈ I
4 |   interact with A through I_n; collect A(I_n)
5 |   S ← Retrain(S ∪ (I_n, A(I_n)))
6 end
  /* S is now a constructed surrogate of A. Analyze S                  */
7 return Analysis(evidence) is Acceptable /* Return false if some violations are found */
```

As an illustrating example, consider the same flight search platform, driven by algorithm A. Let us in this case assume the target platform does not declare anything on the techniques A relies on to rank flights $F(c) = f_0, \ldots, f_j$ (maybe merely using the term "relevance"). A typical Sherlockaudit task would be to study and understand how A ranks its results. Hence, an approach here could consist in collecting many example rankings $F(c_1), F(c_2), \ldots$, and study without any prior different factors (cost, but also duration, number of layovers, departure time, company) that could explain (correlate) with the observed rank of flights.

A typical use case for such a task would be to show that A deliberately favors the flights of some company they are in business with. This example relates to the historical case of SABRE, American Airline's flight reservation system, that used "screen science" to favor its own flights over its competitors by systematically presenting competitors on the second page of the search results [31].

With Sherlock, and as opposed to Bobby, the input data can be fully crafted. This means that there is no prerequisite for a dataset; data can be forged with the objective of triggering some specific behavior for the remote algorithm. This is precisely what line 3 in Algorithm 2 builds on: new input/output pairs are used to retrain a surrogate, which outputs will become more and more similar to the ones of the audited algorithm.

2.4 Reduction to Sherlock

We now list a set of notorious research works that fall into the Sherlockaudit form.

Surge price forecasting for Uber

We refer to the paper entitled "Peeking Beneath the Hood of Uber" [12], where authors rely on some data capture (measured supply, demand, estimated waiting times and surge prices) to fit three linear regression models. Their aim is to predict the surge multiplier in the next 5-min interval. The inputs are crafted from using several smartphones, for in particular bringing a variety of locations in the inputs to the algorithm.

This fits directly the core of Algorithm 2, where a surrogate is trained from the queried data, so that after the query budget is over, the surrogate is used to perform a final test.

Tracking action consequences in outputs with XRay

The audit by Lécuyer et al. [8] creates fake accounts to make them interact with the audited platform (Gmail for instance), and detect which data input (*e.g.*, an email) has likely triggered a particular output (*e.g.*, a received ad in Gmail). Distinct ads on each account are tracked. A correlation engine is run, to associate inputs and outputs. To that end, the placement of inputs on given accounts is crucial to be able to properly infer associations.

A Bayesian model is proposed as a surrogate to simulate the audited service, by computing the probability to observe certain outputs given targeting associations.

Explaining ML decisions with LIME

The goal of LIME [11] is to explain the decisions of a machine learning model in the vicinity of a given input x, by training sparse linear surrogate models as explanations. These models are thus local surrogates. Input samples are drawn uniformly at random around x, to obtain a perturbed dataset (along with its labels returned by the remote model).

COMPAS A celebrated example of `Sherlock`audits is the analysis of COMPAS,[4] an algorithm used by judges, probation and parole officers to assess a criminal defendant's likelihood of becoming a recidivist. This study was used to whistleblow on the bias present in the audited models.

2.4.1 Limits of `Sherlock`

While `Sherlock`audits can in principle target any algorithm, it comes at a price: first, a greater cost, both the number of human intervention required to exploit obtained results and in the amount of requests such results usually require. Second, it leads to a weakened power as the conclusions are ultimately drawn from a model whose interpolation capabilities can always be questioned.

`Sherlock`audit forms first usually require more human intervention in their design and exploitation. This can first be explained from a purely computational perspective: `Bobby`audits are bound to extract a binary information from the target (*i.e.*, L is true or false); hence literally extracting the minimal amount of information, while `Sherlock`audits are supposed to extract much more information. Consider for instance the airline ranking audit: while the `Bobby`information only verifies the statement issued by the platform, `Sherlock`is supposed to come up with a narrative identifying the behavior of the target by generalizing a handful of observations. Typically involved steps are: identify potential functions corresponding to the observed behavior, test and validate potential functions, confirm or infirm each one, confirm conclusions drawn on the surrogate with the real target, and so on. Such steps are time-consuming, and involve a wide variety of skills such as exploratory analysis and statistics.

The second limiting aspect is the number of requests. Indeed, training a local surrogate model naturally exhibits a trade-off between the training set size (*i.e.*, the number of requests issued to the target algorithm) and the accuracy of the resulting surrogate. Hence, to achieve good accuracy, large volumes of input data are often required. As a result, means to automate the generation of inputs to query the target are often necessary. A positive side effect is that generated data will not be protected (unlike personal data). However, such an automation might not be easy, or require considerable human intervention.

3 The law perspective on audits

We now discuss the interplay of audits with the law (as summarized in Fig. 1).

3.1 Legal issues regarding the identity of the auditor

The legal status of the auditor has key consequences on two aspects of the auditing process: (1) the legal risks incurred by the auditor and (2) the judicial opportunities for the audit results.

First, a regulatory authority, commissioned by the law to evaluate the compliance of online algorithms to specific regulations, is granted specific powers, which is removing all legal risks a regular person would face. Likewise, a contract between the auditor and the audited algorithm's operator allows a private auditor to realize some actions which would represent a violation of legal terms for third parties. Researchers do not benefit from specific legal protection yet. Auditing an algorithm without authorization of a national authority nor the consent of the owner (or operator) of the audited algorithms exposes the auditor(s) to lawsuits.

Second, the relationship between the auditor and the audited algorithm also plays an important role in which ways audit results can be used. If it turns out that the audit results of the targeted algorithm highlight illegal content or behavior, this does not allow the auditor to take legal action by himself. Indeed, to start legal action, the law requires a "standing", which is a legal term referring to the existence of an interest in the claim for the claimant.[5] In other words, the claimant (the auditor in our context) must gain a benefit or avoid a loss through the court action.

[4] COMPAS stands for "Correctional Offender Management Profiling for Alternative Sanctions". About the 2016 analysis: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[5] Legal requirement in article 31 of the French code of civil procedure.

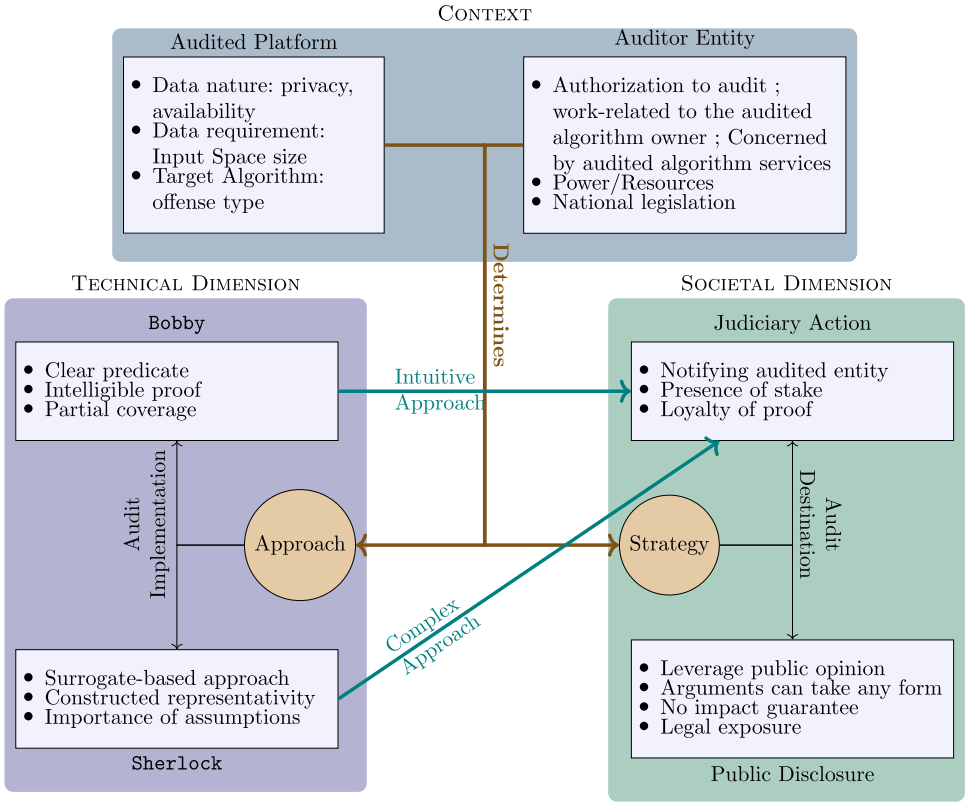

Fig. 1 Overview of the perspectives presented in this article: a given situation (auditor, target algorithm, and offense) defines both possibilities in terms of legal outcomes and in terms of auditing approaches. In favorable situations, law defines a clear predicate that can simply be tested online. When no such predicate exists, auditors have to infer more about the target. Relying on such inferences in a court case is more complex, and hence, those audits are often used to leverage public opinion through whistleblowing

And this is often not the case for most of the research activities. The possibility that remains is then to take the audit results to a national authority or to an association that can act on behalf of the people it represents, having therefore interest to start legal action (see, for instance, actions by consumer associations).

3.2 Legal issues regarding the inputs used to audit

From a legal perspective, a clear split exists based on how the inputs used by the two audit forms are chosen. While `Sherlock`fully crafts its inputs for the purpose of its investigations, `Bobby`uses preexisting data (*i.e.*, existing pictures or user profiles) as a basis for its audit.

Nowadays, existing data of all types can be protected by plenty of legal texts. The GDPR [32] presents the rules of protection on personal data[6]. Every action made on personal data is called *processing*[7] these data. Furthermore, there are special categories of personal data which are subject to additional protections to be processed (health, national security, or even private data matters). Using existing data as inputs for an audit exposes the auditors to the risk of processing personal data, while fully crafted do not. Therefore, `Bobby`can lead to more legal obligations and legal risks[8] for the auditors, whereas `Sherlock`avoids it. Data can also be protected, regarding its economic value, as the content of a database [33], but also as an artistic creation through a copyright law (*e.g.*, as a picture, a drawing). Again, using existing data can lead to process protected data, exposing auditors to legal risks.

These are just two examples of the many regulations applicable to data. From health data to passengers' data, their processing without compliance with legal obligations can expose to important economic sanctions, if not criminal ones in specific situations.

In short, the following counter intuition holds: `Bobby`audits, which look easier to design than `Sherlock`audits, expose the to more legal risks because of the processing of existing data.

3.3 Scientific proof vs legal evidence

3.3.1 Admissibility of the proof

When scientists perform an investigation or an experiment, they produce results that will be evaluated by their peers

[6] GDPR defines *personal data* as "any information relating to an identified or identifiable natural person" which is an extremely wide definition.

[7] Article 4 of the GDPR defines processing as "any operation or set of operations which is performed on personal data", e.g., collection, recording, consultation, alteration, use, etc.

[8] Violation of the GDPR can lead to administrative fines up to 20 million of euros or 4% of the total worldwide annual turnover.

within their research community. The results are evaluated depending on the criteria of the community and this will decide the impact on the field. Nevertheless, the acceptance by a scientific community has no influence on the judge's decision during a case. Law has its own criteria when it comes to the admissibility of legal evidence. It is a specific discipline in law studies called *procedural law*. Depending on the legal system the auditors are operating in, those can be written explicitly or not. To illustrate the notion of "admissibility", we are going to take the example of the French legal system. In France, the admissibility of evidence depends on the loyalty of the establishment of the evidence. The "loyalty of proof" principle is a recurring principle there. However, the specification of this principle will change with the field of law the auditor is working in. As an example, the loyalty of proof principle is different in criminal law and in administrative law, and different in civil law. Even within these legal domains, the principle could differ from one action to another. Law is all about contexts and exceptions.

The legality of evidence In particular, the legality of the evidence principle is described in the French Civil Code[9]. It means that evidence obtained illegally cannot be used in court afterward. For instance, elements found in a computer or any IT system without the proper authorization[10] to do so will not be accepted by the judge, even with proofs of the defendant guilt. Indeed, a piece of evidence obtained by means of an unfair process is inadmissible. Some exceptions to this principle exist in criminal law or in labor law.

Acceptation of evidence The acceptance of the found evidence relies on the interpretation of the judge. The infringement(s) must be necessary and proportionate[11] to the purpose, *i.e.*, finding evidence to support a claim. Necessary evidence means the illegal or disloyal evidence brought by the claimant is their only solution to support their claim. In other words, was the infringement or disloyalty necessary to support the claim? Then, the judge will perform the proportionality test. She will balance the opposed side's rights and liberties that have been violated by the audit establishing the evidence and the evidence stakes for the claimant. This analysis of the necessity and the proportionality is solely based on facts. It means that no generalization can be made out of the decisions given by the judges already[12]. In the end, if the judge qualifies the proof as necessary and proportionate, the evidence will be admitted in court.

Consequences on the choice of the auditing form This concern echoes the inputs used by the two forms of audits and the way the audit was conducted. When someone is auditing without proper authorization, not only does the author get responsible for the violation of rights realized during the process, but also she is making the admissibility of the proof in court more difficult. If the purpose of the audit is to sue a platform, one has to be extremely cautious about the violation realized during this audit. Otherwise, the potential legal consequences of an investigation will be jeopardized. Designing an auditing algorithm that makes the less severe violation of the law is important to increase the chances of admissibility by satisfying (at least theoretically) the condition of necessity. The fact that a better-suited audit algorithm could have been used to search for the targeted element will reduce the chances to see the evidence being admitted in court.

3.3.2 The probative value of the audit results

We have discussed the admissibility of scientific proof in court. But being admissible in the court is not related to the level of importance given to the scientific proof. Legal rules distinguish between the admissibility of the evidence and its evidential value. Admissibility means whether or not the elements will be accepted in the list of evidence, while the value of the proof means the value given in the court by the judge of the evidence that has been accepted beforehand. The probative value of an element can be written explicitly in the law or be let to the sovereign discretion of the judges.

Intrinsic limitation Beforehand, algorithms face a fundamental limit regarding probative value. Algorithmic audits of algorithms target technical elements, written in code, which are an interpretation of a natural language requirement, law, or others, made by the auditors. Translating unclear terms of a legal text enforces an interpretation from the auditor: it is subjective and may not be adequate to the state of law and interpretation of the judge. Audits will always succeed more in identifying simple and explicit illegal situations, which are clearly defined in legal texts or by court decisions. In consequence, the output probative value can always be challenged on the basis that it does not represent what the law definition meant in the first place: this is an inherent limit to all audits to be aware of before taking legal risks, while auditing an algorithm.

[9] Article 9: "Each party has the burden of proving in accordance with the law the facts necessary for the success of its claim."

[10] About the sanctions of fraudulent access in an IT system, see art 323-1 of the French penal code.

[11] These notions of necessity and proportionality of legal evidence have been admitted in the first place by the European Court of Justice. For more information, see J. Van Compernolle, "Les exigences du procès équitable et l'administration des preuves dans le procès civil", RTDH 2012. 429.

[12] G. Lardeux, "Le droit à la preuve: tentative de systématisation", RTD civ. 2017. 1.

The probative value of `Bobby` `Bobby`allows for automatic research through a digital environment which could have been done manually (by a lot of persons tasked to do so): the result can be found again by a person if a confirmation of the existence of a given output is required. There is one positive and one negative legal consequence regarding this. Starting with the negative one, the `Bobby`audit is useful in court only when manifest evidence is needed. By "manifest" evidence, the law refers to illegal content or situation which are noticeable. There is no need for further justification for the claimant in addition to the audit results. The utilization of `Bobby`audits in the goal of legal proceeding is limited by the fact that they can only be found in existing apparent elements in the environment. However, because those elements are existing and are apparent, their correctness is guaranteed; therefore, their probative value should remain equal to the same evidence brought by hand. Using `Bobby`audits should have consequences on the admissibility of the evidence but should not impact its evidential value.

The probative value of `Sherlock` On the other hand, `Sherlock`audits provide interpolations that are most of the time hidden from the user. A working condition for a `Sherlock`audit is to create a representation of the audited algorithm. Because, by definition, the audited algorithm is not open to the public, the correctness of the representation is not guaranteed. Every statement made about the surrogate model is an assumption. As an illustration, we can take the example of someone who tries to assess the potential discrimination bias against women within a hiring decision system [34]. This potential inaccuracy, inherent to all statistical tools, will have consequences regarding the probative value given by the judge to the audit results. This new question has not been addressed by courts yet. Therefore, it is not possible to propose an evidential value to `Sherlock`audits. It could become strong evidence, like DNA in a paternity test[13], or it could be a more contextual, secondary evidence as DNA in criminal proceedings[14].

It is crucial to realize that the use of an audit algorithm, with all the state-of-the-art elements of the scientific procedure followed, does not guarantee that the output will be considered as important evidence from a court, compared to a testimony or other kind of evidence. The consideration of the importance of potential inaccuracy and postulates in audits from a user perspective will take time for lawmakers or judges.

In that light, `Sherlock`audits are good candidates to trigger ethical debates, around findings that may be deemed inappropriate in some societies. When these debates exhibit consensus and a compatible nature, obligatory norms may follow [3]. Independently of norms, ethical misconducts may be reported through whistleblowing [35], which we now discuss.

3.3.3 The whistleblowing alternative

According to the Council of Europe, *"whistleblowing is about bringing into the open information on activities that have harmed or threaten the public interest. People blow the whistle because they believe that these activities should be stopped and remedial action taken"*.

Whistle-blower regulation initiatives emerge in some national law.[15] In 2019, 6 years after the Snowden scandal, the European union voted a directive on the protection of whistle-blowers [36]. The directive grants a legal protection for people who "acquired information on breaches in a work-related context [...][16]. In the situation of audit algorithms, *work-related context* means that the whistle-blower had, has, or is going to have a work relationship with the owner of the controversial audited algorithm. In June 2022, the Council of Europe released a report[17] assessing the impact of the directive. Whistle-blower protection is also part of the public ethics framework adopted by the Council of Europe in 2020.[18]

3.4 The promising role of academics in platform auditing

3.4.1 The "vetted researchers" status

Nowadays, auditing the algorithms of platforms without specific grants exposes the auditor to several legal risks. However, this existing situation might end soon *e.g.*, in Europe with the regulation on a single market for digital (Digital Services Act or DSA for short) brought by the European Commission [37]. The proposal provides guarantees for civil society, including researchers, for the control of platforms. Beyond the creation of a *trusted flaggers* status,

[13] On the importance of DNA in a paternity test, see, among others: Cour de cassation, civile, Chambre civile 1, 25 September 2013, 12−24.588, Inédit, 2013. and Cour de cassation, civile, Chambre civile 1, 25 September 2013, 12−24.588, Inédit, 2013.

[14] Because DNA does not provide all the elements necessary to establish guilt, its usefulness and utilization are actually limited. See Julie Leonhard, " La place de l'ADN dans le procès pénal ", Cahiers Droit, Sciences & Technologies, 9, 2019, 45–56.

[15] In France, a legal protection is granted in 2016 through a law for transparency and against corruption.

[16] Article 4 "*personal scope* of the European directive"

[17] Evaluation report on Recommendation CM/Rec(2014)7 on the protection of whistle-blowers https://www.coe.int/en/web/cdcj/activities/protecting-whistleblowers

[18] Guidelines of the Committee of Ministers of the Council of Europe on public ethics (2020), E.h Section.

the European proposal enforces the "very large online platforms"[19] to provide access to data to *vetted researchers*.

To be qualified as a vetted researcher, a person needs to satisfy four conditions: a) to be "affiliated with academic institutions", b) to be "independent from commercial interests", c) to have and prove their "expertise in the fields related to the risks investigated or related research methodologies", and d) to commit and to be able to "preserve the specific data security and confidentiality". Access to the data will not be asked directly to the very large platforms targeted by the research but through the Digital Services Coordinator, a national authority designated by each European member state.[20]

This vetted researcher's status is announced with the following purpose: the identification and understanding of systemic risks created by very large online platforms. Those systemic risks can be related to the functioning of the platforms or the use of services. Article 26 gives the list of these risks; therefore, the description of the purposes that can be pursued by vetted researchers: a) "dissemination of illegal content", b) "any negative effects for the exercise of the fundamental rights to respect for private and family life, freedom of expression and information, the prohibition of discrimination and the rights of the child", and[21] c) "intentional manipulation [...] with an actual and foreseeable negative effect on the protection of public health, minors, civic discourse, or actual or foreseeable effects related to electoral processes and public security". This new legal status is an important first step forward for the algorithm audit community and a call for the involvement of civil society in digital issues.

3.4.2 Limitations of the vetted researchers' status

The European Council gave its final approval to the Regulation of the DSA in October 2022, but months can pass before its application. The vetted researchers is a status created to legally allow researchers, mostly academics, to identify and understand the systemic risks existing in very large online platforms. First, the mentioned systemic risks are not limited to very large online platforms: illegal content, discrimination or privacy are issues that can also be found in smaller online platforms. Second, these risks are not limited to online platforms at all and can exist in offline applications as well. Researchers may still perform audits before the DSA is applied, or in a context that is out of the vetted researchers' scope of action. In these situations, the guidelines described in this article are still to consider.

[19] Term defined by article 25 of the DSA as "online platforms which provide their services to a number of average monthly active recipients of the service in the Union equal or higher than 45 million [...]."

[20] Article 38 of the Digital Services Act.

[21] Rights from articles 7, 11, 21, and 24 of the Charter of Fundamental Rights of the European Union.

4 Conclusion

We have discussed that the outcomes of audits—that we fit into two categories (`Bobby`and `Sherlock`)—do not necessarily conforms to what is a proper building of a case. Multiple precautions must be carried out before and during the audit, in order for evidence to be considered, so that they have an impact in practice. A central objective also is to avoid the auditor to be prosecuted.

There are an increasing number of approaches to auditing algorithms. In particular, there are some variations of `Bobby`that use real inputs to create synthetic ones, which will in turn be used against the algorithm [40]. This makes the boundary between `Bobby`and `Sherlock`more difficult to distinguish. As an auditor, it is necessary to continuously monitor the development of auditing techniques and evolving laws.

Finally, we stress that the absence of proof of bias in an audited algorithm does not give any certainty. This echoes the so-called *diesel gate*, and more technically the possible temptation for *fairwashing* [9], where the audited algorithm can "sandbox" the auditor into an acceptable vision of its operation. In such a case, more advanced regulation is needed for a proper impact on decision-making algorithms.

Declarations

Conflict of interest The authors declare there are no competing interests with this work.

References

1. Diakopoulos, N.: Accountability in algorithmic decision making. Commun. ACM **59**(2), 56–62 (2016). https://doi.org/10.1145/2844110
2. Ledford, H.: Millions of black people affected by racial bias in health-care algorithms. Nature **574**(7780), 608–610 (2019)
3. Carrillo, M.R.: Artificial intelligence: From ethics to law. Telecommunications Policy **44**(6), 101937 (2020)
4. Klonick, K.: Content moderation modulation. Commun. ACM **64**(1), 29–31 (2020). https://doi.org/10.1145/3436247
5. Metcalf, J., Moss, E., Watkins, E.A., Singh, R., Elish, M.C.: Algorithmic impact assessments and accountability: The co-construction of impacts. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21, pp. 735–746. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445935
6. Mökander, J., Floridi, L.: Operationalising ai governance through ethics-based auditing: an industry case study. AI and Ethics **3**(2), 451–468 (2023)
7. Moore, E.F.: Gedanken-experiments on sequential machines, 129–154 (2016)

8. Lécuyer, M., Ducoffe, G., Lan, F., Papancea, A., Petsios, T., Spahn, R., Chaintreau, A., Geambasu, R.: Xray: Enhancing the web's transparency with differential correlation. In: 23rd USENIX Security Symposium (USENIX Security 14), pp. 49–64. USENIX Association, San Diego, CA (2014). https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/lecuyer
9. Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., Tapp, A.: Fairwashing: the risk of rationalization. In: International Conference on Machine Learning, pp. 161–170 (2019)
10. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency, pp. 77–91 (2018)
11. Ribeiro, M.T., Singh, S., Guestrin, C.: "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 (2016)
12. Chen, L., Mislove, A., Wilson, C.: Peeking beneath the hood of uber. In: Proceedings of the 2015 Internet Measurement Conference, pp. 495–508 (2015)
13. Ribeiro, M.H., Ottoni, R., West, R., Almeida, V.A.F., Meira, W.: Auditing radicalization pathways on youtube. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20, pp. 131–141. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3351095.3372879
14. Bandy, J., Diakopoulos, N.: Auditing news curation systems: a case study examining algorithmic and editorial logic in apple news. Proc. Int. AAAI Conf. Web Soc. Media **14**(1), 36–47 (2020)
15. Galdon Clavell, G., Martín Zamorano, M., Castillo, C., Smith, O., Matic, A.: Auditing algorithms: On lessons learned and the risks of data minimization. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. AIES '20, pp. 265–271. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3375627.3375852
16. Panigutti, C., Perotti, A., Panisson, A., Bajardi, P., Pedreschi, D.: Fairlens: auditing black-box clinical decision support systems. Inf. Process. Manag. **58**(5), 102657 (2021). https://doi.org/10.1016/j.ipm.2021.102657
17. Huszár, F., Ktena, S.I., O'Brien, C., Belli, L., Schlaikjer, A., Hardt, M.: Algorithmic amplification of politics on twitter. arXiv:2110.11010 (2021)
18. Kaiser, J., Rauchfleisch, A.: The implications of venturing down the rabbit hole. Int. Policy Rev. **8**(2), 1–22 (2019)
19. Petropoulos, G.: A European union approach to regulating big tech. Commun. ACM **64**(8), 24–26 (2021). https://doi.org/10.1145/3469104
20. Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., Eling, M., Goodloe, A., Gupta, J., Hart, C., et al.: Governing ai safety through independent audits. Nat. Mach. Intell. **3**(7), 566–571 (2021)
21. UNESCO, C.: Recommendation on the ethics of artificial intelligence. UNESCO France (2021)
22. Jobin, A., Ienca, M., Vayena, E.: The global landscape of ai ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019)
23. Raji, I.D., Xu, P., Honigsberg, C., Ho, D.: Outsider oversight: Designing a third party audit ecosystem for ai governance. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 557–571 (2022)
24. Mökander, J., Schuett, J., Kirk, H.R., Floridi, L.: Auditing large language models: a three-layered approach. AI and Ethics, 1–31 (2023)
25. Matte, C., Bielova, N., Santos, C.: Do cookie banners respect my choice? Measuring legal compliance of banners from IAB Europe's transparency and consent framework. In: 2020 IEEE Symposium on Security and Privacy (SP), pp. 791–809 (2020). IEEE
26. Le Merrer, E., Trédan, G.: Remote explainability faces the bouncer problem. Nat Mach Intell **2**, 529–539 (2020). https://doi.org/10.1038/s42256-020-0216-z
27. Adi, Y., Baum, C., Cisse, M., Pinkas, B., Keshet, J.: Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In: 27th USENIX Security Symposium (USENIX Security 18), pp. 1615–1631. USENIX Association, Baltimore, MD (2018). https://www.usenix.org/conference/usenixsecurity18/presentation/adi
28. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259–268 (2015)
29. Urman, A., Makhortykh, M., Ulloa, R.: Auditing source diversity bias in video search results using virtual agents. In: Companion Proceedings of the Web Conference 2021, pp. 232–236 (2021)
30. Maho, T., Furon, T., Le Merrer, E.: Surfree: a fast surrogate-free black-box attack. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10430–10439 (2021)
31. Sandvig, C., Hamilton, K., Karahalios, K., Langbort, C.: Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and discrimination: converting critical concerns into productive inquiry **22**, 4349–4357 (2014)
32. Réglement (UE) 2016/679 du Parlement Européen et du Conseil du 27 Avril 2016 Relatif à la Protection des Personnes Physiques à L'égard du Traitement des Données à Caractére Personnel et à la Libre Circulation de Ces Données, et Abrogeant la Directive 95/46/CE (règlement Général sur la Protection des Données) (Texte Présentant de L'intérêt Pour l'EEE). http://data.europa.eu/eli/reg/2016/679/oj/fra Accessed 2020-02-26
33. Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the Legal Protection of Databases. http://data.europa.eu/eli/dir/1996/9/2019-06-06/eng. Accessed 09 May 2022
34. Raghavan, M., Barocas, S., Kleinberg, J.M., Levy, K.: Mitigating bias in algorithmic employment screening: evaluating claims and practices. arXiv:1906.09208 (2019)
35. Boot, E.R.: The Ethics of Whistleblowing. Routledge (2019)
36. Directive (EU) 2019/1937 of the European Parliament and of the Council of 23 October 2019 on the Protection of Persons Who Report Breaches of Union Law. Code Number: 305. http://data.europa.eu/eli/dir/2019/1937/oj/eng. Accessed 02 Dec 2021
37. Commission Européenne: Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts com/2021/206 final. In: 2021/0106 (2020)
38. Johnson, D.G., Verdicchio, M.: Ethical ai is not about ai. Commun. ACM **66**(2), 32–34 (2023). https://doi.org/10.1145/3576932
39. Brown, S., Davidovic, J., Hasan, A.: The algorithm audit: Scoring the algorithms that score us. Big Data & Soc 8(1) (2021). https://doi.org/10.1177/2053951720983865
40. Mahmood, K., Mahmood, R., Rathbun, E., Dijk, M.: Back in black: a comparative evaluation of recent state-of-the-art black-box attacks. arXiv:2109.15031 (2021)