Ada Lovelace Institute

DKUK

Report focus
This report clarifies terms in algorithm audits and algorithmic impact assessments, and the current state of research and practice.

Examining the Black Box

Tools for assessing algorithmic systems

Contents

Key takeaways

As algorithmic systems become more critical to decision making across many parts of society, there is increasing interest in how they can be scrutinised and assessed for societal impact, and regulatory and normative compliance.

Clarifying terms and approaches

Through literature review and conversations with experts from a range of disciplines, we've identified four prominent approaches to assessing algorithms that are often referred to by just two terms: algorithm audit and algorithmic impact assessment. But there is not always agreement on what these terms mean among different communities: social scientists, computer scientists, policymakers and the general public have different interpretations and frames of reference.

While there is broad enthusiasm amongst policymakers for algorithm audits and impact assessments there is often lack of detail about the approaches being discussed. This stems both from the confusion of terms, but also from the different maturity of the approaches the terms describe.

Clarifying which approach we're referring to, as well as where further research is needed, will help policymakers and practitioners to do the more vital work of building evidence and methodology to take these approaches forward.

Two terms, four approaches

We focus on **algorithm audit** and **algorithmic impact assessment.** For each, we identify two key approaches the terms can be interpreted as:

- **Algorithm audit**
  - **Bias audit:** a targeted, non-comprehensive approach focused on assessing algorithmic systems for bias.
  - **Regulatory inspection:** a broad approach, focused on an algorithmic system's compliance with regulation or norms, necessitating a number of different tools and methods; typically performed by regulators or auditing professionals.

- **Algorithmic impact assessment**
  - **Algorithmic risk assessment:** assessing possible societal impacts of an algorithmic system *before* the system is in use (with ongoing monitoring often advised).
  - **Algorithmic impact evaluation:** assessing possible societal impacts of an algorithmic system on the users or population it affects *after* it is in use.

For policymakers and practitioners, it may be disappointing to see that many of these approaches are not 'ready to roll out'; that the evidence base and best-practice approaches are still being developed. However, this creates a valuable opportunity to contribute – through case studies, transparent reporting and further research – to the future of assessing algorithmic systems.

Snapshot: tools for assessing algorithmic systems

Table

	Algorithm audits		Algorithmic impact assessments	
	Bias Audit	**Regulatory inspection**	**Algorithmic risk assessment**	**Algorithmic impact evaluation**
What?	A targeted approach focused on assessing algorithmic systems for bias	A broad approach focussed on an algorithmic system's compliance with regulation or norms, and requiring a number of different tools and methods	Assessing possible societal impacts of an algorithmic system before the system is in use (with ongoing monitoring advised)	Assessing possible societal impacts of an algorithmic system on the users or population it affects after it is in use
When?	After deployment	After deployment, potentially ongoing	Before deployment, potentially ongoing	After deployment
Who by?	Researchers, investigative journalists, data scientists	Regulators, auditing and compliance professionals	Creators or commissioners of the algorithmic system	Researchers, policymakers
Origin	Social science audit studies	Regulatory auditing in other fields e.g. financial audits	Environmental impact assessments, data protection impact assessments	Policy impact assessments, which typically are evaluative after the fact
Case study	'Gender shades' study of bias in classification by facial recognition APIs	UK Information Commissioner's Office AI auditing framework draft guidance	Canadian Government's algorithmic impact assessment	Stanford's 'Impact evaluation of a predictive risk modeling tool for Allegheny County's Child Welfare Office'
Status	More established methodology in algorithm context; limited scope	Emerging methodology, skills and capacity requirements for regulators, more established approaches for compliance teams in tech sector	Some established methodologies in other fields, new to algorithm context; requiring evidence as to its applicability and best practice	Established methodology new to algorithm context; requiring evidence as to its applicability and best practice

Introduction

We rely on algorithmic systems for more, and higher stakes, decision making across society: from content moderation and public benefit provision, to public transport and offender sentencing. As we do so, we need to know that algorithmic systems are doing the 'right thing': that they behave as we expect, that they are fair and do not unlawfully discriminate, that they are consistent with regulation, and that they are furthering, not hindering, societal good. In order to understand possible impacts of algorithmic systems and improve public trust in them, there is also an increased imperative for transparency, accountability and oversight of these systems. As algorithms augment, assist and eventually replace human-mediated processes, we need to have confidence in them, to understand the impact they are having and be able to identify their harmful, unlawful or socially unacceptable outcomes.

These challenges from the public, media, policymakers, developers, product managers and civil society, give rise to the question: how can algorithms be assessed? In 2016, the Obama Whitehouse 'big data' report called for the promotion of 'academic research and industry development of algorithmic auditing and external testing of big data systems to ensure that people are being treated fairly... [including] through the emerging field of algorithmic systems accountability, where stakeholders and designers of technology "investigate normatively significant instances of discrimination involving computer algorithms" and use nascent tools and approaches to proactively avoid discrimination...'[1] Since that time, policy and technical discussions have been circling around options or means for assessment, most commonly the 'algorithm audit' or 'algorithmic impact assessment'. But there is not always agreement on what these terms mean amongst different communities: social scientists, computer scientists, policymakers and the general public have different interpretations and frames of reference.

In synthesising research and policy documents related to algorithm assessment tools, this report breaks down the most commonly discussed terms and assigns them to the range of approaches that they can describe. Each of these approaches have different merits and contexts in which they may be helpful. The goal of clarifying these terms is to move past confusion, create shared understanding and focus on the important work of developing and evaluating different approaches to algorithmic assessment.

This report is primarily aimed at policymakers, to inform more accurate and focused policy conversations. It may also be helpful to anyone who creates or interacts with an algorithmic system and wants to know what methods or approaches exist to assess and evaluate that system.

Two methodologies: audit and impact assessment

There are two methodologies that have seen wide reference in popular, academic, policy and industry discourse around the use of data and algorithms in decision making: **algorithm audit** and **algorithmic impact assessment.**[2]

These terms have been used variously and interchangeably by, for example, pioneering mathematician Cathy O'Neil, who called for algorithm audits in *Weapons of Math Destruction*;[3] the UK Information Commissioner's Office, which is developing an algorithm auditing framework;[4] and the AI Now Institute, whose recommendation of establishing algorithmic impact assessments was followed by the Canadian Government. In 2019, the German Data Ethics Commission made algorithmic risk assessments a policy recommendation.[5] Meanwhile, the field of fairness, accountability and transparency in machine learning has grown, yielding practical processes to mitigate the potential harms (and maximise the benefits) of algorithmic systems.

These different perspectives mean that, while the two terms are increasingly popular, their meanings can vary. Here, we unpack the approaches and possible interpretations alongside further research and practice priorities.

Algorithm Audits

Algorithm audits have become a handy catch-all for panellists and policymakers responding to demands for, or advocating for, more accountability around the use of algorithmic systems, in particular those that underscore the large tech platforms. In the UK, the Centre for Data Ethics & Innovation has even recognised a growing market opportunity for the UK to be at the forefront of 'an AI audit market', capitalising on increasing interest in audit as a mechanism for assessment, accountability and public trust and confidence.[6] However, through literature review and conversations with experts, we find the term 'audit' is used by different actors in different ways.

We surmise that confusion about algorithm audits comes from the two relevant meanings of 'audit':

1. **Audit from the perspective of the computer science community**, which proposes adopting the social science practice of an audit study and applying it to algorithmic systems. This form of audit is a narrowly targeted test of a particular hypothesis about a system by looking at its inputs and outputs – for instance, seeing if it has racial bias in the outcomes of a decision. In this paper, this is called a **bias audit**.
2. **Audit from the perspective of its use in common language** to mean a broad inspection and compliance exercise, such as a financial audit. In this sense audit is being used to describe a comprehensive inspection to check if an algorithmic system is behaving according to rules or norms. In this report, this is called a **regulatory inspection**.

Making these distinctions between algorithm audits and algorithm inspections allows us to focus on using the right approaches in the right contexts, and the important work of developing best practice in each form of audit.

Both types of audit above refer to practices that can be potentially used to assess algorithmic systems as a means of external (arms-length, independent) accountability, such as that sought by civil society, regulators or the media. They are also both processes that the creators of algorithmic systems may wish to emulate internally to verify whether such systems will withstand external scrutiny, and pre-empt possible problems with a system. This may be done by the organisation commissioning an external party, or through conducting internal bias audits or inspections themselves. For instance, a company building an AI hiring system may run a bias audit against its own system to look for discrimination against people displaying protected characteristics. Similarly, some have suggested a range of inspection tools that could be applied internally,[7] perhaps pre-empting concerning findings by a regulatory inspection.

Bias audit

Typically this form of audit is conducted by external, independent actors who are completely outside of – and don't enjoy the collaboration of – the team or organisation designing and deploying the algorithmic system. Bias audits aren't 'audits' in the sense of financial audits, which attempt to comprehensively check every part of a system using a range of qualitative and technical measures. Instead, a bias audit is a narrowly targeted test of a particular aspect of a system – for instance, seeing if it has racial bias in the outcomes of a decision. This type of approach builds on social science 'audit studies',[8] which are field experiments in which researchers test for forms of discrimination in social processes by participating in them: for instance, sending identical job applications with different names and looking at the results according to perceived gender or ethnicity of the names.[9]

Bias audits are usually done on algorithmic systems already in use, typically by people not involved in the development of the system. As a result, they generally don't look at the code of the system. Instead, they compare the data that goes into the system with the results that come out. They are therefore sometimes referred to as 'black box testing' or 'black box audits' as they treat the system as a black box, only looking at the inputs and outputs of the system.

The exact techniques used for bias audits will vary depending on the system, its purpose, the context of its use and access to its inputs, outputs or algorithms. However, work on auditing for discrimination in online platforms is particularly developed, with Sandvig et al. laying out a range of research methods and approaches to implementing them in different contexts:[10]

- **Scraping audit:** the researcher writes a program to make a series of requests to a website or API for an algorithmic system and observe the results. Challenges include risk of breaching a platform's terms of service and the US Computer Fraud and Abuse Act (CFAA),[11] however recent rulings in district court have made it clear scraping and activities to probe algorithmic systems for discrimination that breach the terms of service are not in violation of the CFAA.[12]

- **Sock puppet audit:** where a classic audit study might have involved hiring actors, or creating fake CVs, a sock puppet audit creates fake user accounts to observe the operation of the system.[13]

- **Crowdsourced/collaborative audit:** the researcher recruits users to perform the test; the same as a sock puppet audit but with real human users instead of fake accounts.[14] A current example is *Who Targets Me*, in which volunteers add a browser extension to monitor the political advertising they are being shown, and thereby crowdsource information about political ad targeting.[15]

The majority of published bias audits have been conducted by independent researchers or investigative journalists. However, bias audit techniques can also be applied by the developers of a system to their own system. This may be done in-house, or by commissioning a third party, and would provide more access to the system than a typical external bias audit. There are limitations, however, both in the level of accountability and challenge that may come without independence, and in the lack of supported capacity for this in most tech firms.

While techniques might differ, the uniting feature of bias audits is that they require a concrete hypothesis: a particular metric or feature that is being tested for. These metrics are usually classifications of humans – race, gender, age etc. – similar to those protected characteristics established in antidiscrimination legislation. They are typically socially constructed, and may vary between nations and social contexts, even if the tech that is being analysed transcends them.[16] In using these classifications, researchers are often resorting to legal or scientific definitions that are in themselves contested, flawed or constructed in the context of a biased system and may overlook new axes of discrimination that can occur in algorithmic systems.[17] In addition there are few standard benchmarks for what 'bias' is, to measure against,[18] and – where such benchmarks exist – they may fail to capture the contextual nature of discrimination that investigations of bias seek to tackle.[19] Together, this means bias audits cannot give a holistic picture of the system; a bias audit showing that a system doesn't treat people differently by gender does not mean the system is free of other forms of discrimination issues, or that it might not have other issues or impacts on society to be aware of.

Case study: Gender Shades Algorithm Audit

In 'Gender Shades', Buolamwini and Gebru audited commercial facial recognition APIs to assess their performance at classifying faces by binary gender and to determine if there were accuracy disparities based on gender or race.[20]

This audit was conducted against three commercial recognition APIs: Microsoft, IBM and Face++. Researchers used a dataset containing photographs of people with a wide range of skin types labelled by gender. They then ran these images against each API and recorded whether the API's classification of gender matched the gender label they had. They analysed these results by gender, Fitzpatrick Skin Type, and the intersection of the two. They found that darker-skinned females were the most misclassified group, with a significant disparity in the accuracy of gender classification by gender and skin type.

A year after the initial study's release, Raji and Buolamwini looked into the impact of the 'Gender shades' audit.[21] They re-ran the original audit and found that all target systems of the original audit had released new API versions with reduced accuracy disparities. In particular, the dark-skinned female subgroup saw a 17.7–30.4% reduction in error rate across the systems. Raji and Buolamwini highlight two dimensions they consider influential on the ability of the audit to incentivise the creators of the audited systems to improve them:

- Anonymous vs non-anonymous – revealing the exact system tested may increase public pressure to correct issues identified in auditing.
- Single vs multi-target – performing the same audit on multiple commercial algorithmic systems may enable competitive forces to stronger incentivise response.[22]

Who does or might want to do bias audits?

- **Researchers:** to build evidence around how algorithmic systems behave. For example, researchers have audited Twitter's search algorithm for political bias in search results.[23]
- **Investigative journalists:** to uncover problems with algorithmic systems that are in the public interest. For instance, investigative journalists at ProPublica conducted an external audit of the COMPAS recidivism prediction tool discovering and reporting on racial bias in the system.[24]
- **Civil society organisations:** to investigate algorithmic systems that might affect people they work with or advocate on behalf of. For instance, in the UK the Joint Council for the Welfare of Immigrants has launched a legal case with Foxglove Legal to force the investigation of Home Office visa application algorithms to establish if they are racially discriminatory.[25]

Future research and practice priorities

There's a varied and growing academic literature of bias audits: from auditing social media search results for political bias,[26] to advertising targeting,[27] to content personalisation systems[28] and beyond. When an audit finds a disparate impact, the auditors typically hope to see change in the audited system, and perhaps in other similar systems, or the development practices that created the system. For instance, Raji and Buolamwini examined the impact of publicly naming and disclosing bias in performance AI systems through looking at the commercial impact of the 'Gender shades' audit of facial recognition APIs.[29]

To progress this further, researchers and research funders could consider prioritising:

1. **Developing the meta-literature:** on impact of bias audit work, methods and publishing approaches, with more meta-studies into the effect audits have on the systems they audited. This work should aim to address the question of 'how to have impact with a bias audit that has found disparity?'.
2. **Audits in more contexts:** much of the earlier bias audit literature focused on online contexts – search, social media, advertising and targeting – but the growing work in public sector use cases, commercial APIs and novel scenarios will expand understanding of techniques and approaches. In addition, there are new technical

contexts to consider: establishing good bias auditing methods for complex systems, such as deep reinforcement learning models, where it is harder to interpret the relationship between inputs and outputs.

3. **Audits over time:** most bias audits are conducted once or twice. Algorithmic systems running in the real world are frequently updated, have datasets that change over time and are increasingly using dynamic models. There is a need for more bias auditing approaches that can be conducted in an ongoing, or regular, way.
4. **Funding capacity and influence:** for externally conducted bias audits, researchers, investigative journalists and civil society organisations generally rely on external funding to advance or pursue such research projects. This funding dynamic can pose ethical challenges, and potentially direct the attention or scope of bias audit practice. At the same time, there appear to be insufficient incentives currently for companies to sufficiently resource internal bias auditing. These are unsolved challenges, and tie in with questions about where bias audits and antidiscrimination legislation intersect and when bias audit techniques ought to form part of regulatory inspection, with powers and capacity in regulatory bodies.

Regulatory inspection

A bias audit is able to test the output of a system by deploying certain inputs, but stops short of scrutinising the full lifecycle of a system. A method for inspection of an entire algorithmic system against particular regulations would be better described as regulatory inspection (and might include, but not be limited to, bias audits). A regulatory inspection could be used to assess whether an algorithmic system complied with data protection law, equalities legislation, or insurance industry requirements, for instance.[30]

This type of 'full-service' inspection would need the participation or cooperation of those deploying the algorithmic system. As a result, it is most likely to be conducted by regulators with statutory powers to conduct such inspections, or an auditing professional working with the developers of the system to ensure compliance. For instance, the UK Centre for Data Ethics & Innovation report on online targeting recommends that the UK Government's new online harms regulator should have 'information gathering powers', including 'the power to give independent experts secure access to platform data to undertake audits'.

In practice, this regulatory inspection may apply to an entire product, a model, or an algorithm, depending on sector or usage context. To be robust, however, a regulatory inspection should not be limited to examining code (which is both controversial, and offers a limited and slow understanding of large systems), inputs, outputs and documentation, but also consider an algorithmic system in the context it operates – the organisational processes and human behaviour around it.

While there is a range of internal regulatory inspection practice for compliance within tech companies, there is not a developed methodology for a regulatory algorithm inspection by regulators. and it is difficult to imagine a standardised approach given how context dependent such inspection is. It is likely that sector-specific understandings of regulatory inspections of algorithmic systems are required, and that the scope and functions of regulatory inspections would differ dramatically in vastly different contexts: for example, social media content moderation algorithms and high-frequency trading algorithms. The tools deployed by an inspector might include applying techniques from bias auditing, but also could involve mandating access to data about the algorithm's users, inspecting how the system is operating, speaking with developers or users, or looking at code underpinning an algorithmic system. In practice, while there are growing calls for these processes and the regulatory powers to conduct them, there aren't yet many examples of this in action.

Case Study: UK Information Commissioner's Office Auditing Framework

In the UK, the Information Commissioner's Office (ICO), is developing an auditing framework for AI to inform its inspection of algorithmic systems, particularly with respect to data protection, as well as to inform internal inspections carried out for compliance.[31] This is referred to as 'auditing', meant in the sense of a comprehensive suite of tools for compliance professionals to inspect whether an algorithmic system is complying with data protection obligations.

The draft guidance considers how people might assess and mitigate:

- Accountability and governance.
- Fair, lawful and transparent processing, including system performance, assessment and discrimination mitigation.

- Data minimisation and security.
- Upholding individual rights and freedoms.[32]

The ICO's auditing framework is illustrative of how many methodologies a regulatory inspection might employ. It advises using a range of techniques: identifying and assessing trade offs, bias auditing, explanation and training, and documentation of decision making including legal, organisational, technical and security considerations. It is specifically designed to pertain to the European data protection regime, which adopts a risk-based approach to data protection. It will thus differ substantially from a regulatory inspection developed in other sectors where rule-based approaches to regulation are prominent.

Who does or might want to conduct a regulatory inspection?

- **Regulators:** to assess and investigate potential non-compliance.
- **Auditing professionals:** to ensure organisations' compliance with sector-specific or technology-specific regulation, or broader frameworks such as equality legislation.

Future research and practice priorities

Many regulators and other audit bodies worldwide have not previously had to engage with the idea of algorithm inspection. As policymakers contemplate expanding the remit of regulators to include algorithm inspection, there are numerous gaps to address in both the available legal remit and powers to conduct inspections, and organisational capacity and skill set.

This role is increasingly crucial; for regulators in many areas to have sufficient oversight over the impact of algorithmic systems, they will need to have the knowledge, skills and approaches to thoroughly inspect algorithmic systems and scrutinise how they function, both technically, and within the relevant social context. Further research is needed to understand:

1. What legal powers do regulators need and how should they be defined, either generically or sectorally, in order to appropriately equip regulators with a mandate to develop algorithm inspections and to give public and private sector entities foreseeability about how their systems will be inspected? This includes legal powers concerning auditability by design, compelled disclosure and enforcement.
2. What skills and capabilities do regulators need, and how can these best be developed and shared?
3. What mechanisms are in place to enable regulators to share both successes and failures in developing and using inspection tool suites, to facilitate learning and improvement?

Algorithmic impact assessments

Algorithmic impact assessment can mean different things depending on where in the lifecycle of an algorithmic system they occur, the types of impact and the types of system being assessed. Our research reveals that two interpretations of impact assessment are in use:

1. **Algorithmic risk assessments,** which are used in advance of a system or feature being deployed, in order to assess the possible areas of impact of the system and the attendant risk. This type of methodology is well developed in the context of environmental impact assessments, data protection impact assessments and other forms of risk assessment focused on potential harms.
2. **Algorithmic impact evaluations,** which are conducted after a system has been deployed, and focus on the effects of that system on a particular population. These tend to mirror policy or economic impact assessments.

The respective fields of risk assessments and impact evaluations as a whole are well established, however the research and practice applying these to algorithmic risk and impact is, as yet, niche, with only a small body of work. There are also outstanding questions as to the applicability and efficacy of these approaches in the development, governance and accountability of algorithmic systems.

Algorithmic risk assessments

Algorithmic risk assessments are designed to enable those involved in the creation or procurement of an algorithmic system to evaluate and address the potential impacts of the system. They generally seek to be holistic, looking beyond just the data or model itself, to how it will be used in practice and how users and the wider public will interact with or be affected by it. To date, they have primarily been deployed by, or considered in the context of, the public sector.[33] Impact risk assessments are intended for use before the system is 'live' in the real world, but can also be integrated as a continuous process to monitor changing risks or assess new features. Because they are internal processes, they include scrutiny of non-public details of the system.

Algorithmic impact assessments involve the study of an algorithmic system, begun in advance of deployment, to identify risks and concerns, and to propose means of mitigating those risks and concerns. This approach originates in other forms of impact assessment used in the context of environment regulation, human rights standards and data protection law, which are often legally mandated.

Algorithmic risk assessments generally go beyond considerations of privacy or individual data protection, to wider societal considerations.[34] This has been the approach that led to the introduction of algorithmic impact assessments in Canada, where the 'Directive on Automated Decision-Making' requires Assistant Deputy Ministers responsible for programmes using 'automated decision systems' to conduct an algorithmic impact assessment.[35] In this case, an algorithmic impact assessment means an online questionnaire that works to establish the level of risk of the system, and, depending on the result, will generate further requirements of those responsible for the system. The factors considered include the motivations of the project, stakes of decisions, vulnerability of service users and the type of technology in use.[36] The directive came into force in April 2020.[37]

Calls for algorithmic risk assessments are being used to encourage best practice on the part of government bodies or other organisations deploying algorithmic systems. The AI Now Institute has proposed a process for algorithmic impact assessments intended for public sector agencies 'to assess automated decision systems and to ensure public accountability'.[38] This framework suggests a series of steps that could be undertaken prior to a public sector deployment of a system to form an algorithmic impact assessment, as well as recommending continuing these processes after the system is in use. Experts in the UK have argued that a proper construction of data protection impact assessment obligations under GDPR requires them to go beyond narrow considerations of privacy, reflecting calls for more holistic algorithmic risk assessment models.[39] Similarly, Understanding artificial intelligence ethics and safety, guidance produced for the UK government by the Alan Turing Institute's Public Policy programme, outlines a framework for 'stakeholder impact assessments' that consider all the people that may be impacted by such a system, in order to 'bring to light unseen risks that threaten to affect individuals and the public good'.[40]

There is a wide variety in how algorithmic impact assessments are discussed and required, and little consensus on or evidence of what best practice in this field looks like. There are important questions about whether, and how, they work as mechanisms to affect change, and how accountability and transparency are ensured – both in obligation to follow up on the recommendations of algorithmic risk assessments, and in publishing them for external scrutiny.

Case Study: the RAMSES project impact assessments

The RAMSES project is a collaboration between eleven EU research and policing institutions to build software to help identify and investigate financial cybercrime. It uses web scraping, image, video and data analysis to track the flow of data from malware software and money from malware payments. Its stated aims are to better understand how and where malware is spread and identify the source of these financial cybercrimes.[41]

Trilateral Research, one of the eleven partners, conducted impact assessments to try to incorporate a "privacy-by-design approach during the technology development and a consideration of data ethics to create a proportionate tool for related law enforcement activities".[42] These impact assessments were referred to as a "Privacy and Ethics Impact Assessment", for which the ethics impact assessment fits the general model of an algorithmic impact assessment as it:

- "studies a particular technology, product or service and/or data processing activity;
- identifies risks and concerns;
- proposes means to address and mitigate them."

Who might want to do algorithmic risk assessments?

- **Creators, deployers or procurers of algorithmic systems:** to understand and mitigate possible risks or negative impacts and consider societal implications of their work.
- **Policymakers:** might consider making them a statutory requirement for public or private sector bodies.
- **Public sector organisations:** to build public trust and confidence.

Further research and practice priorities

There has been great policy attention and excitement around algorithmic risk assessments as a means of allaying public concerns about the impact of algorithmic systems. However, there is a lack of standardised approaches or evidence to establish that they work in practice as a governance framework for algorithmic systems.

Further work is needed from researchers and users of algorithmic risk assessments:

1. **Case studies of existing methods in practice:** for instance, there are no published case studies recording the deployment of the

Canadian Algorithm Impact assessment or the AI Now process in the field. Research should document and evaluate how these tools changed or shaped practice and outcomes, enabling the evaluation of their effectiveness as a tool, and informing discussion about how they could be improved.

2. **Learning from algorithmic risk assessments in other fields:** what can we learn from environmental, human rights, data protection and similar impact assessments? Understanding how such mechanisms work in practice, when they are effective and what makes them so will be useful to establish if they are a useful governance mechanism for AI.

Algorithmic impact evaluation

Algorithmic impact evaluation looks at the impact of an algorithmic system on a population, after the system is already in use. This approach stems from traditional policy or economic impact assessments that look, post-hoc, at the impact of new policies, processes or events. Impact evaluations can be conducted by independent researchers, though may need some access to data from the system, such as details of people subject to the system.

Algorithmic impact evaluation appears particularly pertinent in the public sector, where in some cases it is becoming increasingly hard to differentiate policy impact from the algorithmic systems that might be part of the implementation of that policy. While in the private sector there may be a challenge in having sufficient evidence on the population or society prior to the introduction of the system, algorithmic impact evaluation is theoretically applicable across sectors. Algorithmic impact evaluations may draw from the 'Constructive Technology Assessment' from the field of Science and Technology Studies which looks at the wider processes, ecosystem and culture that algorithmic systems are deployed and the multi-directional impact – both of systems on population, but also of population and context on the system.[43]

Human rights impact assessments are also typically conducted post-hoc, and have been both directly adopted within the tech sector, and used as inspiration for proposals of new assessment methods to examine the impact of algorithmic systems. They look at the adverse effects of business projects or activities on rights-holders and their enjoyment of human rights. However, there are questions about accountability – whether developers of these systems have sufficient obligation to enact the recommendations of these evaluations. Facebook, for instance, has taken actions that contradict the recommendations of the human rights impact assessment it commissioned on its systems in Myanmar.[44]

Case Study: Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office

In 2016, Allegheny County in Pennsylvania, USA, introduced predictive risk modelling to their children's welfare office.[45] The Allegheny Family Screening Tool (AFST) presented referral call screeners in children's protective services with a risk score for the children involved to contribute to the decision on whether to further investigate the referral (screen-in) or not (screen-out).

In 2018, researchers at Stanford conducted an impact evaluation comparing outcomes for children involved in children's protective services after the full implementation of the predictive risk modelling tool, to outcomes for children involved in protective services in the period before the system was implemented. They looked at accuracy, case workload, disparities and consistency of outcomes.

They found that implementing the AFST and surrounding policy resulted in 'moderate improvements in accuracy of screen-ins with small decreases in the accuracy in screen-outs, a halt in the downward trend in pre-implementation screen-ins for investigation, no large or consistent differences across race/ethnic or age-specific subgroups in these outcomes, and no large or substantial differences in consistency across call screeners'. They also point out that there could be further work to investigate how these impacts relate to the core goals of child protection, such as safety and children's wellbeing.[46]

There are, however, multiple, sometimes conflicting reviews of the AFST tool.[47] There are also concerns that the Stanford impact evaluation provided legitimacy to these practices, leading to further application of algorithmic decision making in children's social services which raise a complex range of ethical and professional issues.

Who might want to do algorithmic impact evaluations?

- **Public sector and policymakers:** to understand the impact of policies that involve or are often implemented alongside algorithmic systems.
- **Researchers:** to build evidence on how algorithmic systems affect people, communities and society.

Further research and practice priorities

To make algorithmic systems that work for people and society, there must be understanding of what their impact on people and society is over time. Rigorous approaches from social science have clear potential to help understand how these systems are changing outcomes. This is of particular importance to the public sector, but would also be welcome in private sector deployments, particularly as the lines between the two are often blurred in the development and deployment of algorithmic systems. The key next steps are:

1. **Additional published post-hoc impact assessments:** requiring both the research funding, but also the willingness of those developing and procuring these systems to be open to independent research and publishing. This leaves open risk of conflicts of interest in research practice, necessitating opportunities for further independent review and challenge of both the evaluations and systems they evaluate.
2. **Best practice** that clarifies additional skills or considerations for applying these forms of impact evaluation to cases with algorithmic systems and establishing mechanisms to encourage cooperation with recommendations.

Further research and practice priorities

Table

Key stakeholders	Further research and practice priorities
Regulators/auditors	**Focussing on regulatory inspection of algorithms:** - For your sector, or generically, consider what legal powers may be missing to enable regulatory inspection of algorithmic systems; this could include legal powers concerning auditability by design, compelled disclosure and enforcement. - Consider what skills and capabilities you would need to perform this regulatory function, and how these could best be developed and shared. - Share both successes and failures in developing and using inspection tool suites, to facilitate learning and improvement.
Civic society organisations and nonprofits	**Focussing on bias audits:** - Continue pursuing and publishing bias audits of algorithmic systems, including the methodologies used. - Consider analysing, or collaborating with researchers to analyse, the impact of the bias audit approach used and how it may have resulted in change (including sharing failures).
Public sector	**Focussing on algorithmic risk assessment:** - Publish case studies of algorithmic risk assessments conducted, documenting how the process changed or shaped the design, development and outcomes. - Open up to independent researchers and civil society collaboration to help conduct or evaluate this work. **Focussing on algorithmic impact evaluation:** - Additional published post-hoc impact evaluations: requiring both the research funding, but also the willingness of those developing and procuring these systems to be open to independent research. Note that all forms of assessment discussed in this paper could have relevance for public sector organisations or teams deploying algorithmic systems.

Key stakeholders	Further research and practice priorities
Private sector	**Focussing on algorithmic risk assessment:** • Publish case studies of algorithmic risk assessments conducted, documenting how the process changed shaped the design, development and outcomes. • Open to independent researchers to evaluate this work. Note that all forms of assessment discussed in this paper could have relevance for the private sector. An overall consideration is the value of publishing findings from these processes, and enabling access for regulators, researchers, civil society organisations or the public to conduct them.
Researchers	**Meta-evaluative work of these approaches:** • A common theme in the research agenda across these practices is scope for work that evaluates whether these approaches are useful for the governance of algorithmic systems, and, if so, how to design and use them most effectively.
Data scientists and engineers	**Design and develop with audit and assessment in mind:** • Collaborate with others to conduct algorithm risk assessments and impact evaluations. • Consider and grow best practice for designing, documenting and developing systems for bias audit and regulatory inspection. • More technical tools, libraries and frameworks will likely be needed as these methods and practices develop; there may be a range of opportunities in offering the technical skills to work in collaboration with regulators, civil society, researchers and the public sector.

This report and the research behind it were developed in collaboration by the Ada Lovelace Institute and DataKind UK.

At the Ada Lovelace Institute, we are exploring further work on algorithmic assessments, in particular considering the research, policy and practice development required for regulatory inspection. If you are working on projects relevant to this research agenda or topics discussed in this paper, please feel free to get in touch via the details below.

Many thanks to contributors and readers, including: Christine Henry (working with DataKind UK), Madeleine Elish, Francine Bennett, Rashida Richardson, Amba Kak, Andrew Strait, Swee Leng Harris, Lisa Whiting and Reuben Binns.

The Ada Lovelace Institute is a research institute and deliberative body dedicated to ensuring that data and AI work for people and society. Our core belief is that the benefits of data and AI must be justly and equitably distributed, and must enhance individual and social wellbeing.

The Ada Lovelace Institute was established by the Nuffield Foundation in early 2018, in collaboration with the Alan Turing Institute, the Royal Society, the British Academy, the Royal Statistical Society, the Wellcome Trust, Luminate, techUK and the Nuffield Council on Bioethics.

We are funded by the Nuffield Foundation, an independent charitable trust with a mission to advance social well-being. The Foundation funds research that informs social policy, primarily in education, welfare and justice. It also provides opportunities for young people to develop skills and confidence in STEM and research. In addition to the Ada Lovelace Institute, the Foundation is also the founder and co-funder of the Nuffield Council on Bioethics and the Nuffield Family Justice Observatory.

We are named after visionary computing pioneer Ada Lovelace (1815–52), who set high standards for intellectual rigour and analysis in her work and writings, responding to Charles Babbage's Analytical Engine. These qualities, combined with her impressive abilities to see beyond accepted models, aggregate meanings from disparate sources and work with others to build new knowledge, are embedded in our daily work and embodied in the Institute that proudly bears her name.

DataKind UK is a charity with a mission to transform the impact of social change organisations through the use of data and data science. Our focus is on building the capacity of the social sector to use data effectively and responsibly. All our projects are carried out by volunteers, largely pro-bono data scientists working in industry.

Endnotes

[1] Executive Office of the President of the USA. (2016). Big data: a report on algorithmic systems, opportunity, and civil rights. *Archives.gov.* Available at: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf [Accessed 22.4.20].

[2] The term 'algorithmic impact assessment' is used as the term is broadly adopted. Semantically it differs from the naming of other forms of impact assessment (e.g. 'environmental', 'privacy', 'human rights') in that it refers to the impact of *algorithms*, as opposed to the impact on them. We use 'algorithm audit' over 'algorithmic audit' because we are referring to the auditing of the algorithms themselves, rather than using algorithmic means to audit something.

[3] O'Neil, C. (2016). *Weapons of Math Destruction*. Penguin, UK edition.

[4] Binns, R. and Gallo, V. (2019). An overview of the auditing framework for artificial intelligence and its core components. *UK Information Commissioner's Office*. Available at: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-an-overview-of-the-auditing-framework-for-artificial-intelligence-and-its-core-components/ Accessed [22.4.20].

[5] 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 22.4.20].

[6] Centre for Data Ethics and Innovation. (2020). Online targeting: final report and recommendations. *Gov.uk*. Available at: www.gov.uk/government/publications/cdei-review-of-online-targeting Accessed [22.4.20].

[7] Raji, D., Smart, A. et al. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: *Conference on Fairness, Accountability, and Transparency*, p33–44. [online] Barcelona: ACM. Available at: https://doi.org/10.1145/3351095.3372873 Accessed [22.4.20].

[8] Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: research methods for detecting discrimination on internet platforms. *Pre-conference on Data and Discrimination at the 64th annual meeting of the International Communication Association*, p1–23. Available at: http://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf [Accessed 22.4.20].

[9] Bertrand, M. and Mullainathan, S. (2004). Are Emily And Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, v94, p991–1013. Available at: https://scholar.harvard.edu/files/sendhil/files/are_emily_and_greg_more_employable_than_lakisha_and_jamal.pdf [Accessed 22.4.20].

[10] Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: research methods for detecting discrimination on internet platforms. *Pre-conference on Data and Discrimination at the 64th annual meeting of the International Communication Association*, p1–23. Available at: http://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf [Accessed 22.4.20].

[11] Ibid.

[12] United States District Court for the District of Columbia. (2020). Sandvig v. Barr – memorandum opinion. ACLU. [online] Available at: www.aclu.org/sandvig-v-barr-memorandum-opinion ; Williams, J. (2018). D.C. court: accessing public information is not a computer crime. *Electronic Frontier Foundation*. [online] Available at: www.eff.org/deeplinks/2018/04/dc-court-accessing-public-information-not-computer-crime [All accessed 23.4.20].

[13] Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: research methods for detecting discrimination on internet platforms. *Pre-conference on Data and Discrimination at the 64th annual meeting of the International Communication Association*, p1–23. Available at: http://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf [Accessed 22.4.20].

[14] Ibid.

[15] Who Targets Me? Homepage. [online] Available at: https://whotargets.me/ [Accessed 22.4.20].

[16] Sánchez-Monedero, J., Dencik, L., Edwards, L. (2020). What does it mean to 'solve' the problem of discrimination in hiring?: social, technical and legal perspectives from the UK on automated hiring systems. In: *Conference on Fairness, Accountability, and Transparency*, p33–44. [online] Barcelona: ACM. Available at: https://doi.org/10.1145/3351095.3372849 [Accessed 22.4.20].

[17] Mittelstadt, B. (2017). From individual to group privacy in big data analytics. *Philosophy & Technology*, 30, 475–494. Available at: https://doi.org/10.1007/s13347-017-0253-7 [Accessed 22.4.20].

[18] For instance, NIST in the US has benchmarks for bias in facial recognition systems, but not yet other systems such as natural language processing systems: Grother, P., Ngan, M., Hanaoka, K. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. *National Institute of Standards Technology, US Department of Commerce*. Available at: https://doi.org/10.6028/NIST.IR.8280 [Accessed 22.4.20].

[19] Wachter, S., Mittelstadt., B., Russel, C. (2020). Why Fairness cannot be automated: bridging the gap between EU non-discrimination law and AI. [online] Available at: http://dx.doi.org/10.2139/ssrn.3547922 [Accessed 22.4.20].

[20] Buolamwini, J. and Gebru, T. (2018). Gender shades: intersectional accuracy disparities in commercial gender classification. In: *Conference on Fairness, Accountability, and Transparency*, 81, p1–15. [online] New York: PLMR. Available at: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf Accessed [22.4.20].

[21] Raji, I., Buolamwini, J. (2019). Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. In: *Conference on Artificial Intelligence, Ethics, and Society (AIES)*. [online] Honolulu: ACM. Available at: https://dl.acm.org/doi/10.1145/3306618.3314244 Accessed [22.4.20].

[22] Ibid.

[23] Kulshrestha, J et al. (2017). Quantifying search bias: investigating sources of bias for political searches in social media. In: *Conference on Computer-Supported Cooperative Work and Social Computing*. [online] Portland: ACM. Available at: https://dl.acm.org/doi/10.1145/2998181.2998321 Accessed [22.4.20].

[24] Larson, J., Mattu, S., Kirchner, L. and Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica*. [online] Available at: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm [Accessed 22.4.20].

[25] Youles, X. (2019). JCWI to challenge the Home Office over the use of intelligence algorithms. *Immigration News*. [online] Available at: https://immigrationnews.co.uk/jcwi-to-challenge-the-home-office-over-the-use-of-intelligence-algorithms/ [Accessed 22.4.20].

[26] Kulshrestha, J et al. (2017). Quantifying search bias: investigating sources of bias for political searches in social media. In: *Conference on Computer-Supported Cooperative Work and Social Computing*. [online] Portland: ACM. Available at: https://dl.acm.org/doi/10.1145/2998181.2998321 Accessed [22.4.20].

[27] Lécuyer, M. et al. (2014). XRay: enhancing the web's transparency with differential correlation. *USENIX Security Symposium*. [online] San Diego: ACM. Available at: https://dl.acm.org/doi/10.5555/2671225.2671229 [Accessed 22.4.20].

[28] Mittelstadt, B. (2016). Auditing for transparency in content personalization systems. *International Journal of Communication*, 10, p12. Available at: https://ijoc.org/index.php/ijoc/article/view/6267 [Accessed 22.4.20].

[29] Raji, I., Buolamwini, J. (2019). Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. In: *Conference on Artificial Intelligence, Ethics, and Society (AIES)*. [online] Honolulu: ACM. Available at: https://dl.acm.org/doi/10.1145/3306618.3314244 Accessed [22.4.20].

[30] Centre for Data Ethics & Innovation. (2020). Online targeting: final report and recommendations. *Gov.uk*. Available at: www.gov.uk/government/publications/cdei-review-of-online-targeting [Accessed 22.4.20].

[31] UK Information Commissioner's Office. (2020). Guidance on the AI auditing framework. *Ico.org.uk* Available at: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf [Accessed 22.4.20].

[32] Ibid.

[33] Reisman, D. Schultz, J. Crawford, K. Whittaker, M. (2018). Algorithmic impact assessments: a practical framework for public agency accountability. AI Now Institute. Available at: https://ainowinstitute.org/aiareport2018.pdf [Accessed 22.4.20].

[34] While Kaminski and Malgieri consider 'model algorithmic impact assessments' as a means to build on data protection impact assessments, most literature, as discussed further in this paper, presents visions of algorithmic risk assessments that are broader than GDPR concerns. For instance, proposed methods include review by external researchers, encourage public participation and include a right to challenge beyond the scope of existing DPIAs: Kaminski, E. and Malgieri G. (2019) Algorithmic impact assessments under the GDPR: producing multi-layered

explanations. University of Colorado Law Legal Studies Research Paper, 19–28. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3456224; Panel for the Future of Science and Technology (2019). A governance framework for algorithmic accountability and transparency. European Parliamentary Research Service. Available at: www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf [All accessed 21.4.20].

[35] Government of Canada. (2019). Directive on automated decision-making. *Gc.ca*. Available at: www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592 [Accessed 22.4.20].

[36] Government of Canada. (2019). Algorithmic impact assessment. *Canada-cs.github.io*, v0.7.5. Available at: https://canada-ca.github.io/aia-eia-js/ [Accessed 22.4.20].

[37] Government of Canada. (2019). Directive on automated decision-making. Gc.ca. Available at: www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592 [Accessed 22.4.20].

[38] Reisman, D. Schultz, J. Crawford, K. Whittaker, M. (2018). Algorithmic impact assessments: a practical framework for public agency accountability. *AI Now Institute*. Available at: https://ainowinstitute.org/aiareport2018.pdf [Accessed 22.4.20].

[39] Harris, S. L. (2020). Data protection impact assessments as rule of law governance mechanisms. Data & Policy, 2. Available at: https://doi.org/10.1017/dap.2020.3 [Accessed 22.4.20].

[40] Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. *The Alan Turing Institute*. Available at: www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf [Accessed 22.4.20].

[41] RAMSES. (2020). Ramses Project. [online] Available at: https://ramses2020.eu/project [Accessed 22.4.20].

[42] Trilateral Research. (2020). Assessing the privacy and ethical impacts of new technologies: the RAMSES project case study. [online] Available at: www.trilateralresearch.com/assessing-privacy-ethical-impacts-new-technologies-ramses-project-case-study/ [Accessed 22.4.20].

[43] Konrad, K., Rip, A. and Schulze Greiving, V. (2017). Constructive technology assessment – STS for and with technology actors. *European Association for the Study of Science and Technology Review*, 36(3). Available at: https://easst.net/article/constructive-technology-assessment-sts-for-and-with-technology-actors/ [Accessed 22.4.20].

[44] Warofka, A. (2018). An independent assessment of the human rights impact of Facebook in Myanmar. Facebook. Available at: https://about.fb.com/news/2018/11/myanmar-hria/ ; Wong, J. C. (2019). 'Overreacting to failure': Facebook's new Myanmar strategy baffles local activists. *The Guardian*. [online] Available at: www.theguardian.com/technology/2019/feb/07/facebook-myanmar-genocide-violence-hate-speech [All accessed 22.4.20].

[45] Goldhaber-Fiebert, J. and Prince, L. (2019). *Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office*. Stanford University and Allegheny County Department of Human Services.

[46] Goldhaber-Fiebert, J. and Prince, L. (2019). *Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office*. Stanford University and Allegheny County Department of Human Services.

[47] For example: major concerns raised with the system by Eubanks, V. (2018). *Automating Inequality*. St. Martin's Press, New York; limitations of measures of fairness and challenges of practical implementation acknowledged in Chouldechova et al. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In: Conference on Fairness, Accountability, and Transparency, 81, p1–15. [online] New York: PLMR. Available at: http://proceedings.mlr.press/v81/chouldechova18a/chouldechova18a.pdf [Accessed 19.4.20].

adalovelaceinstitute.org
@AdaLovelaceInst
hello@adalovelaceinstitute.org

Ada Lovelace Institute
28 Bedford Square
London WC1B 3JS
+44 (0) 20 7631 0566
Registered charity 206601

DataKind UK
datakind.org.uk
@DataKindUK
contact@datakind.org.uk
Charity no. 1154213
Supported by Open Society Foundations

Designed by Soapbox, www.soapbox.co.uk