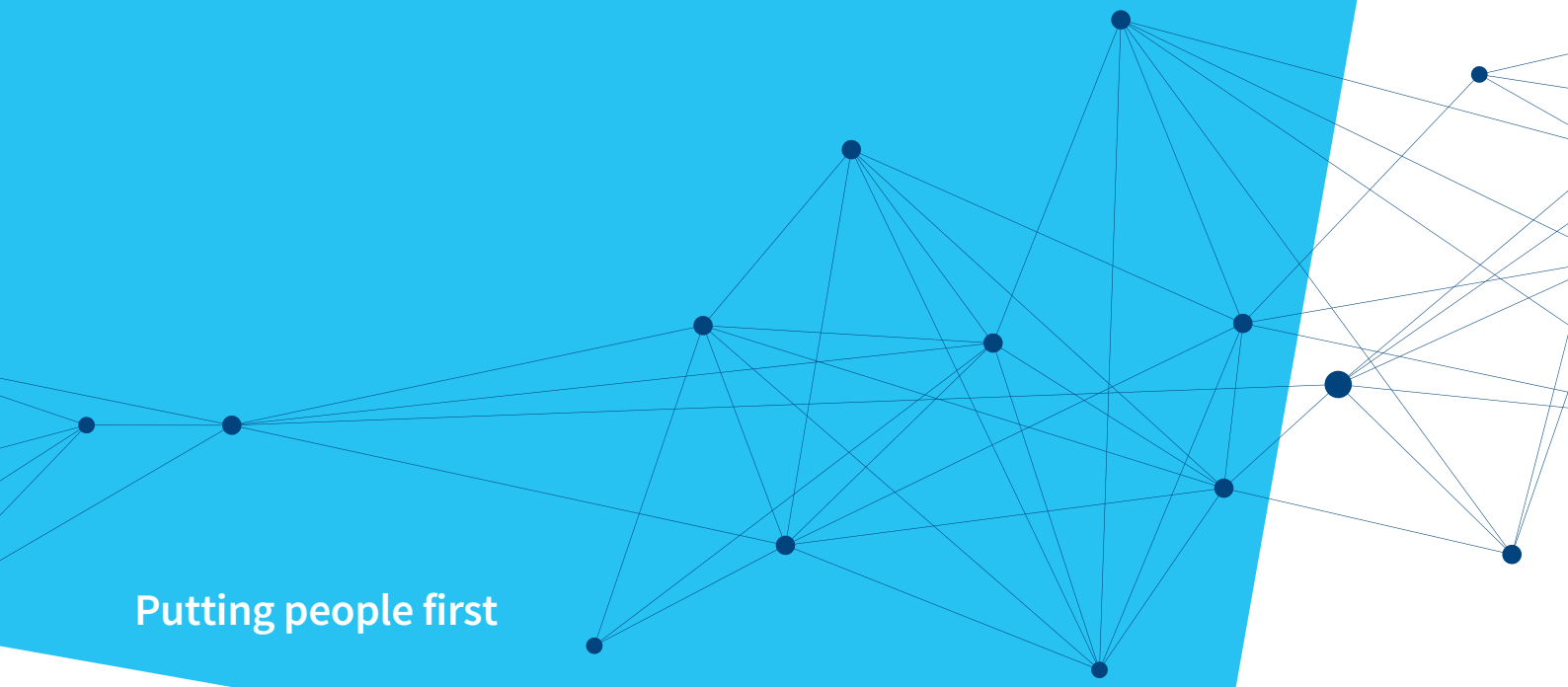
Institute for the Future of Work

Policy Briefing

Building a systematic framework of accountability for algorithmic decision making



Putting people first

Contents

Written by
Stephanie Sheir
Dora Meredith
Dr Abigail Gilbert
Anna Thomas

With thanks to
Helen Mountfield QC
Dr David Leslie
Dr Logan Graham
Dr Emre Kazim
Associate Professor Reuben Binns
Sa'ad Hossain QC

Published
November 2021

Executive Summary

The National AI Strategy recognises that the impact of AI on the UK and the wider world will be profound over the next decade[1]. Algorithmic systems are increasingly used by Governments and businesses to make 'automated', data-driven decisions that are having far-reaching consequences on work and working people, business and communities. Existing regimes for regulation have been outpaced, requiring new proposals to ensure meaningful accountability, safeguard fundamental values and shape innovation in the public interest.

This policy paper proposes algorithmic impact assessments as a systematic framework of accountability for algorithmic decision making to support the next phase of the AI Strategy. The UK Government recognises the need to build on new algorithmic transparency standards in the public sector,[2] and is considering the merits of a unified impact assessment as a form of assurance for AI[3]. This briefing analyses existing models and case studies in the light of new evidence about use of algorithmic systems at work, highlighting key components to inform new regulation.

New evidence suggests that the UK proposal should:

- Establish a new corporate duty to undertake, disclose and act on pre-emptive Algorithmic Impact Assessments (AIAs) in the public interest.
- Ensure that this duty applies from the earliest stage of designing algorithmic systems.
- Mandate rigorous ex ante assessment and ongoing evaluation of risks, impacts and anticipated impacts through deployment. This requires compliance with four essential planks for AIAs.
- Forefront assessment of impacts on equality and good work on people whose interests are likely to be affected.
- Enable future development of context-specific protocols, guidance and standardised techniques and methods of evaluation, covering use of algorithmic systems at work.

The Institute for the Future of Work focuses on the frontier of changes to work but our insights and recommendations are relevant to the wider debate on AI governance and regulation. At a global level, the UNESCO Global Recommendations on the Ethics of AI in November 2021 forefronts the principles of human dignity, inclusive growth and social justice. This invites particular attention to the impacts of AI on work – and provides a normative basis for our framework of accountability.

Introduction

Interest in algorithmic risk and impact assessments is increasing, as evidence is growing on the wide-ranging impacts that algorithmic systems may have, including upon access to, terms and conditions, and quality of work.[4] This briefing introduces algorithmic impact assessments as a systematic framework of accountability for algorithmic decision-making and decision support systems. Based on analysis of existing tools and case studies in the light of new research,[5] we propose a regulatory framework that integrates individual and systemic approaches to accountability. Our proposal could shape the forthcoming AI White Paper, extend the remit of the Online Harms Bill or—as we propose is best—trigger a new, overarching Accountability for Algorithms Act.[6] It also informs the Government's current consultation on the GDPR: Data, a new direction.[7]

This briefing builds on recent national and international work on algorithmic accountability from the Ethics Team at the Alan Turing Institute,[8] Ada Lovelace Institute workshops, the CDEI's Bias Review[9] and APPG on the Future of Work's inquiry into the New Frontier.[10] Draft amendments for potential legislation (V1) are attached at Annex 1, based on the insights in this paper.

The Institute for the Future of Work focuses on the changing world of work. Work is the thread that connects people's everyday experience with their communities, the economy and the state, public policy and private ventures. Addressing the challenges and opportunities of algorithmic systems at work is key in its own right, but a deep-dive into the work context acts as a lens to understand the implications of automated decision-making in multiple high stakes environments. Recent emphasis in AI Ethics on the social and economic impacts of algorithmic systems highlight the particular significance of considering the consequences for work and working people, suggesting policy and legislation must specifically consider this context.

Our goal is to inform legislation to shape innovation in the public interest, safeguard our fundamental values and bring accountability to the algorithmic systems increasingly shaping people's lives.

What is an Algorithmic Impact Assessment?

An algorithmic impact assessment ('AIA') should be seen as an overarching approach, as well as a procedure for risk management and accountability. An AIA focuses on key decision-making points in the design and deployment of an algorithmic system, requiring careful assessment of risks and impacts before real-world use. This ensures these socio-technical systems are human-centred and accountable, reducing risks and maximising benefits for people and society as a whole. AIAs are conceptualised and practiced very differently by companies, academics and policy makers, with the concept used to describe a range of accountability tools and processes across the spectrum of soft to hard governance. An AIA should integrate the technical, human, organisational and social elements of evaluation over time. AIAs can, and should, be applied to advance defined goals, such as the principle that AI should promote people's wellbeing.

Why do we need a systematic framework of accountability for algorithmic decision-making?

The introduction of an AIA framework across the technology lifecycle could provide a unified approach to accountability for algorithmic decision making and a standardised mechanism to enable performance of the assessment in practice. Evidence in the work context demonstrates that technical approaches commonly deployed by technology companies before deployment of algorithmic systems are inadequate.[11] Socio-technical approaches are needed which recognise and make explicit the human decisions that take place both before deployment, and within institutional contexts from conception of the system. AIAs have traditionally been used to identify harms but

there is increasing interest in their broader application, from shaping the design of an automated system to enabling action in response to an impact assessment.

In the UK, policymakers are committed to a regulatory framework that offers meaningful accountability for algorithmic systems deployed in both private and public sectors and are considering AIAs as an instrument to deliver this.[12] The framework should be guided by overarching principles and duties, supported by detailed guidance from the regulators, and at a sectoral level. This approach recognises a range of real and potential impacts that may be widely allocated and not immediately ascertainable, as new research in the work context demonstrates.[13] An AIA framework should allow different methods and measures for their evaluation, as well as use of the full range of emerging tools for the governance of the AI ecosystem, including bias audits, regulatory inspection, voluntary algorithmic impact assessments and public transparency initiatives.[14] It can also combine the best of existing impact assessments, including those for data protection, equality and human rights.

With clear goals and essential standards established for evaluation, the AIA should enable articulation, better measurement and stronger accountability for such a range of impacts; shape design to take these impacts into account from conception through the technology life cycle; improve multi-layered transparency and explanation through record-keeping; enable context-appropriate response in ways which apply established principles; and allow for ongoing evaluation and improved iterations as practice, guidance and case law builds.

Given the relative infancy of existing initiatives globally,[15] the UK is well-placed to combine strengths in innovation and governance and lead in this crucial policy field to shape the global conversation on AI regulation and governance.

Insights on Algorithmic Impact Assessments

Our review points to the following insights on building a systematic regulatory framework for accountability grounded in algorithmic impact assessments:

- Overarching principles should offer direction and normative baselines for evaluation by AIAs. This should be combined with procedural and substantive duties.
- Primary legislation should specify essential requirements needed to fulfil the new AIA duties but not how these should be fulfilled. Regulation must allow for context-specific response as sectoral guidance is developed to avoid or ameliorate specific harms.
- The AIA framework should span the value chain and innovation cycle, from concept, through design, procurement, deployment and any changes in use. Assessment must be dynamic and rigorous.
- The framework should not focus on the model or on content in any way that might restrict or obscure examination of evaluating a complex, socio-technical system or the practices behind their production, deployment and use.
- Mechanisms to pinpoint actors and stakeholders, and construct a system for ongoing engagement, are necessary to perform accurate ex ante and ex post facto evaluation of real-life impacts. This would 'future proof' the model by allowing for the discovery of new impacts.
- Boosted capacity and resources would enable the DCRF and regulators to investigate, certify and attach conditions to use of algorithmic systems; and to run AIA sandboxes to inform guidance.
- Existing regulatory requirements such as requirements for human rights, equality and data protection impact assessments, should be incorporated within the AIA 'umbrella[16].'
- Impacts on good work need explicit attention in regulation and subsequent guidance.

“

Human beings and organisations that use machines of this kind have to take responsibility. If we don’t design the future we want, the future will be designed by accident.

Helen Mountfield QC, Expert in constitutional, human rights and equality law

Quote taken from the APPG report,
The New Frontier: Artificial Intelligence at Work

Types of AIA models

We have identified three primary models of AIAs: the Questionnaire Model, the Data Protection Impact Assessment (DPIA) model and the Public Agency Model.[17] Here, we describe the models and highlight key challenges and lessons learnt from each one.

Model 1

The Questionnaire Model

The questionnaire model involves an AIA completed in question and answer format for public sector AI applications, and is reflected in the approach taken by Canada's Directive on Automated Decision-Making 2019, stipulating that government agencies must complete an AIA before deployment of an Automated Decision System (ADS). The questions used by the Canadian AIA include motivation behind and the need for automation as well as the degree of explanation and human involvement in the system. A point scoring system is then used to create an assessment of risk level. This model employs questions that are fairly simple and invites short or yes/no answers rather than detailed statements or assessments (see appendices for sample).

Although only three AIAs have been published pursuant to the Directive in Canada, with 12 further assessments expected for publication shortly, the Treasury Board of Canada Secretariat (TBCS) reports that undertaking the process has been helpful to raise awareness, identify challenges and enable a reflective exercise to consider the real purpose and intended outcomes. Making set requirements, without specifying the methodology, works well. Details of the assessment were not always reported to the public when they could be, and this might allay concerns.

Peer review, triggered by higher risk environments, was helpful to deepen the evaluation. Other relevant assessments, such as a privacy assessment for data processing or the Canadian GBA+ -gender based analysis, have been included by adding specific questions to elicit these assessments. Impacts on work are not currently part of the assessment but the TBCS recognises this is a notable gap.

The Canadian questionnaire model is light but some features appear to be working well, in particular the tiered approach to requirements as risk increases; peer review; and incorporation of other assessments required by law. A wider review of impacts is needed.

Model 2

The DPIA Model

Data protection impact assessments (DPIAs), mandated in GDPR for data processing that is likely to result in a high risk to individuals, could serve as a model for an AIA. DPIAs as a governance mechanism are currently under review.[18] Although DPIAs are aimed at protecting people from risks arising as a result of data processing, rather than algorithmic decision making, the focus in the GDPR on automated decision making, impact assessment and procedures for risk mitigation are instructive, noting the high levels of data processing ordinarily involved in Automated Decision Systems (ADS).[19] The overarching principles of GDPR, including those of fairness, transparency, purpose limitation and accountability, offer overall direction. The principle of 'data protection by design' is a novel feature that explicitly requires consideration of data protection principles at concept and design stages. The Information Commissioner's Office (ICO) recognises the need for an update of its employment practices code, which is underway.[20]

Policy Briefing

Disclosure and external review of DPIAs are not required (although some responsible employers have started to disclose DPIAs on a voluntary basis)[21] leading the DPIA model to fall short on public engagement. Without such disclosure and transparency about the fact, purpose and remit of an ADS system, it is difficult for people to understand or engage with relevant decisions.[22] Although Article 35 (9) stipulates that data controllers ought to seek the views of data subjects where appropriate, there is no collective or individualised due process relating to this requirement under GDPR.[23] The limitations of DPIAs have been noted in terms of promoting accountability and ensuring fairness in the work context.[24] and the ICO is producing dedicated guidance to address the challenges of new and pervasive AI applications.[25]

The principle-based framework of the DPIA offers a sound model for AIAs. The principle of 'data protection by design' is instructive. Well-recognised limits of DPIAs, including disclosure, remit and stakeholder engagement could be addressed by legislating for an overarching AIA.

Model 3

The Public Agency Model

The most comprehensive AIA model is the public agency model that formalises relationships of accountability between public agencies to the public and to regulators, through the ongoing exercise of transparency, stakeholder dialogue and regulatory compliance.[26] Its primary elements have been proposed and endorsed by the European Parliamentary Research Service (EPRS).

The public agency model involves a number of ex ante risk assessment procedures of varying levels of stringency. Unlike DPIAs, the public agency model features a great degree of public engagement, both through the disclosure of information, solicitation of feedback and due process challenge period for impacted communities.[27] This approach could be adapted to differing levels of AI risk, types of organisation and accountability requirements, and for the private sector. Research on use of algorithmic systems at work points to the erosion of sharp distinctions between the private and public sectors, given design and procurement of AI systems tend to originate in the private sector.[28] The public agency model has not been applied to work or private applications to date.

The public agency model points to the advantages of establishing an ongoing process for evaluation, alongside wider consultation, that would enable dynamic and responsive monitoring of impacts and harms. The pre-emptive procedures and tiered approach outlined are instructive.

The structure and requirements of most AIA models derive from existing public interest impact assessments (IAs), particularly data protection ('DPIA') and human rights ('HRIA'). These assessments share a number of components designed to measure the impacts of automated decision-making systems ('ADS') against predetermined metrics and to create accountability through the formalisation of relationships between corporations, affected stakeholders and regulatory forums. AIAs can either be required through legislative mandate, solicited voluntarily as a function of reputational or corporate responsibility-based concerns, or in response to proven harms.[29]

Policy Briefing

When should an AIA be triggered?

Precise risk thresholds for the initiation of an AIA are an area of practical concern, as case studies in the work context show[30] and gap in the literature, as scholars make almost no reference to requirements for triggering an AIA.

The proposed Algorithmic Accountability Act 2019[31] in the United States requires AIAs when ADS are implemented which affect certain sensitive domains of people's lives. Similarly, the EU AI Act employs thresholds of risk. Generally, 'high risk' or 'sensitive' processing is said to involve impacts on the legal or other interests of individuals, processing of personal data, or would otherwise have 'significant effects' on a person's life circumstances, especially where vulnerable individuals are involved. This would include impacts on access to work, pay, terms and conditions and quality of work.[32]

The Canadian model requires completion of an AIA if an ADS is used to recommend or make any administrative decisions in the public sector, including when used as support for human decision-makers.[33] Although this classification is broad, the Treasury Board of Canada Secretariat (TBCS) report that, in practice, AIAs are required for ADS which make decisions about individuals, impacting on their rights, access to services or regulatory action.[34] The use of ADS for automating workflows that do not target individuals or corporations are not currently within scope of the Directive, although the TBCS are actively seeking to extend it to require consideration of impacts on work in light of recent evidence on impacts. The TBCS report that all published and ongoing AIAs have been assessed as risk 2/3, with neither low nor high risk cases recognised.

Guidance refining the risk-based approach taken by DPIAs in Article 35 of GDPR offers some inspiration, notwithstanding criticism for requiring too high and uncertain a threshold for assessment, leading to inconsistent application. DPIAs are required when personal data processing is likely to result in a 'high risk' to an individual's fundamental rights and freedoms.[35] The EDB has confirmed this means decisions that 'significantly affect' individuals including decisions that affect the circumstances of the individual for a prolonged period, and financial impacts. This means that decisions about recruitment, pay, work allocation, terms and conditions would all be covered, as IFOW analysis demonstrates.[36] Less obvious uses which impact on job quality include social scoring in recruitment algorithms, systematic monitoring, the use of biometric data and/or location-based tracking, all of which meet definitions of high risk outlined by ICO and Article 29 Working Party guidance.[37]

Fundamentally, the initiation of an AIA must accord with an understanding of how and when an ADS can be appropriately adjusted or its harms mitigated, before significant or irreversible elements are put into place.[38]

Examination of 'triggers' suggest that AIAs should be routinely undertaken in work context and whenever there is a risk of impact on access, fundamental conditions, or the nature of work. Given particular challenges raised by AI, including the scale, reach, and predictive capacity of AI applications,[39] there are strong grounds for requiring an AIA for any use of AI in a work environment. In such sensitive environments, the value of a risk-based threshold is very limited. AIA triggers invite specific attention and guidance from the regulators in any event, once legislators establish an appropriate threshold in principle.

> Companies should be required and encouraged to consider and remedy any adverse impacts as soon as possible in the innovation cycle, not post event. The opportunity is for the UK to do better, and lead globally.

Tabitha Goldstaub, Chair of the UK Government's AI Council

Quote taken from the APPG report,
The New Frontier: Artificial Intelligence at Work

Developing the AIA process

Overall approach

Although the models for AIAs vary in focus and comprehensiveness, AIAs generally conform to a 4-stage process.[40] This section explores these stages in more detail, drawing out insights from applications at work to inform the details of our regulatory model.

Impacts from algorithmic systems are determined by human decisions throughout the entire life-cycle of a product.[41] Some impacts will be better detected before, or after, the product is deployed, as analysis in the work context demonstrates. While impacts on equality can be significantly shaped by human decisions in-product refinement,[42] impacts on broader dimensions of good work—such as learning, dignity and autonomy— may only become visible through choices made during implementation.

Recognising how these human choices determine impacts is critical to overcoming the regulatory lacuna in governance, particularly in the AI context. There is a lack of clarity in popular writings on the character of the AI "black box". On the one hand, this refers to the proprietary protectionism of firms that are attempting to safeguard their intellectual property by not disclosing details about their software and computer code. This intentional lack of transparency is often cast as financially prudent and strategically necessary in competitive innovation environments. But it may also be used to set up unjustifiable roadblocks to sensible regulatory oversight and evade reasonable expectations about public-facing assurance of fair practices.

Figure 1: The 4 stage AIA process

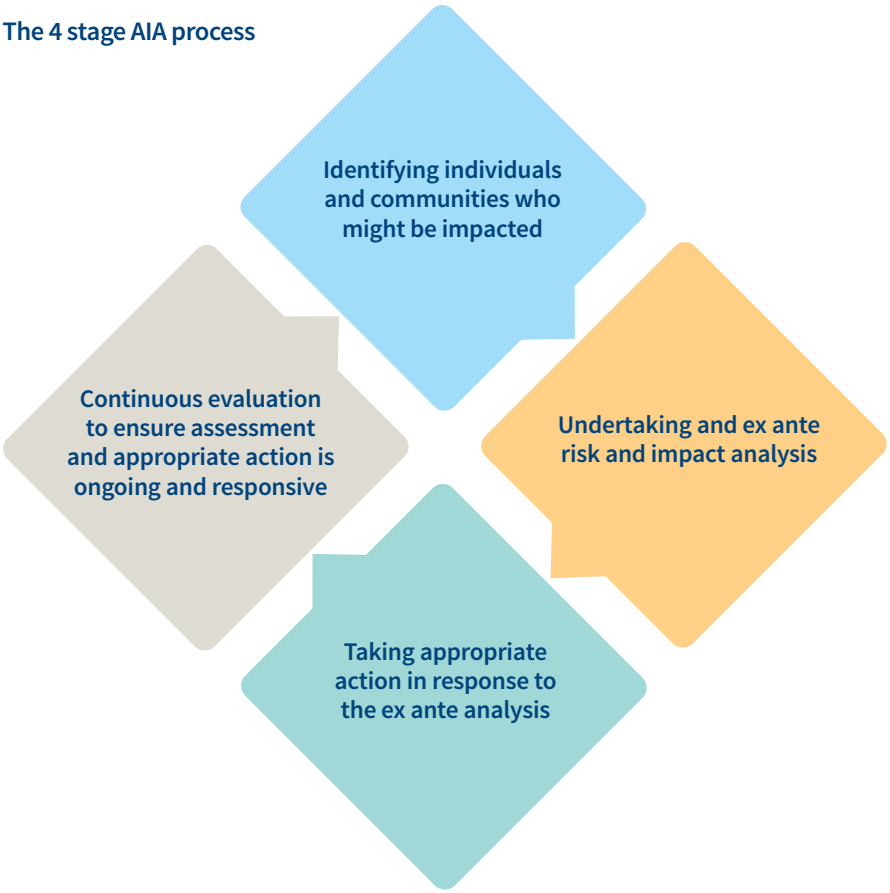
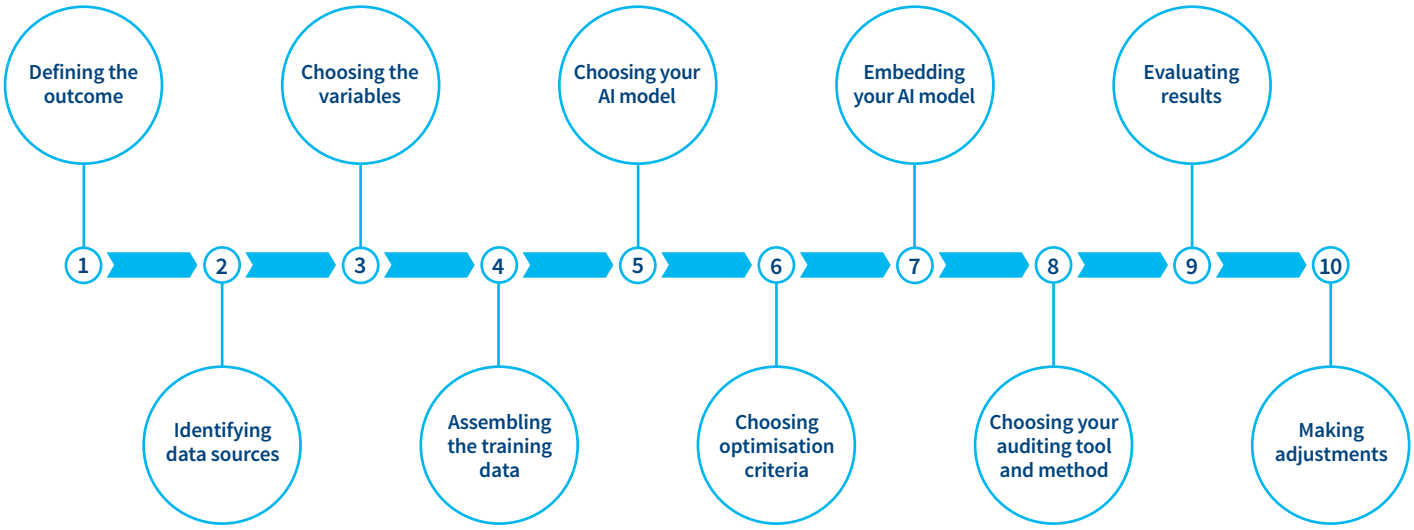


Figure 2: Key decision-making points (with thanks to Dr Logan Graham)



On the other hand, the AI "black box" refers to the barriers to understanding that complex algorithms pose. This kind of opaqueness may lead to the setting up of a different sort of obstacle to regulatory intervention as human agency is ceded to "smarter" machine learning systems. This perspective is not supported by the literature: human decisions determine the design objectives of even the most advanced machine learning tools.[43]

Stage 1

Identifying individuals and communities who might be impacted

Stage 1 of an AIA consists of identifying those who might be impacted by automated decision-making.

Proposals regarding procedures for identifying impacted groups are sparse in the AIA literature.[44] The Alan Turing Institute suggests employing a 'stakeholder impact assessment' which identifies affected stakeholders in the first instance, by considering which groups and individuals are negatively impacted by the ADS, with particular attention to the most vulnerable.[45] The use of 'Human Impact Statements' has also been suggested, involving comparative population analysis of impact.[46] Potentially impacted populations and their status-based categories, such as race and gender, are identified, before the potential impact of uncertainty or statistical error on these sub-populations is assessed.[47] This methodology sketches out how impact on relevant populations ought to be considered but falls short of pinpointing the relevant affected communities.

Policy Briefing

Once relevant impacted populations have been identified, engagement with these communities can take place. Types of community engagement procedures include public comment periods, focus groups, surveys and representative boards.[48] In the literature, community engagement is largely addressed through reference to engagement with the wider public through public disclosure of information, with less reference to engagement with impacted communities, however defined. Typically, communities of interest are framed in broad terms, such as the consumer or citizen, as opposed to workers or specific algorithmic subjects.

Lessons from Human Rights Impact Assessments (HRIAs)

Identification of affected stakeholders has received more treatment in the human rights literature, especially through Human Rights Impact Assessments (HRIAs), which are gaining traction in popularity and attention.[49] In guidance on conducting HRIAs, the UN highlights the methodology of impact zoning, recommended by the World Bank, which seeks to identify individuals affected at each stage of business projects.[50]

Impact zoning involves the following steps:

i) Sketching of key design components of the project that may cause social or environmental impacts. These impacts give rise to 'impact zones'.

ii) Identify and overlay broad stakeholder groups over the impact zones.

iii) Consult with stakeholder representatives and verify which groups are affected by which impacts.

iv) Further consideration of particularly vulnerable groups by identifying the groups which may be disproportionately affected by business activities due to their disadvantaged status.

These guidelines are sparse in detail, probably because identification of affected groups is a highly context-dependent exercise. However, the broad principles of the identification of disadvantaged groups and consultation with stakeholder representatives and are instructive for the design of AIAs.[51]

Identifying the individuals and communities who might be impacted will also form the basis for multi-stakeholder engagement and participation through the process of assessment.[52]

The identification of affected communities, particularly the most vulnerable, represent a gap in the literature. More needs to be done on the design of methodologies to identify the relevant impacted individuals and communities of automated decision-making, with lessons to be drawn from human rights impact procedures.

Policy Briefing

Stage 2

Undertaking an ex ante risk and impact analysis

The second stage of the AIA is a risk and impact analysis undertaken before deployment of the ADS.

IFOW case studies and literature review demonstrates that analysis should start with an overview of the system including its purpose, scope, intended use and potential implementation timeline. In the work context, this has been shown to be necessary to encourage thoughtful use, understand impacts and avoid work places becoming sites of experimentation.[53] A clear definition allows the developer to clarify the preliminary details of an ADS, including which parts of the greater workflow are automated and which are human-controlled, overseen or supervised. In practice, a statement of purpose setting out the capabilities, remit and proposed outcome is more useful to many stakeholders than details of programming, variables or validation, although in-depth assessment will need full access.[54]

Next, the analysis is implemented through either an *internal* self-assessment of the system or an *external* commissioned assessment conducted by a third-party organisation to evaluate potential impacts on stakeholders, such as inaccuracy, bias and other harms.

There is no universal methodology for conducting the risk and impact analysis, although we have identified common elements in proposed AIA frameworks, including the use of internal or external auditing, quantitative and qualitative methodologies, and various measurements of impact. As impact is contextually defined, metrics and thresholds of risk and impact may vary. Public disclosure of information could be hindered by opaque language by entities seeking to fulfil compliance requirements without being fully transparent.[55]

The model of 'datasheets for datasets' promotes transparency through the accompaniment of datasets with information sheets on dataset motivation, composition, intended uses and so forth.[56] To maximise the effectiveness of disclosure of information, so that appropriate courses of action can be undertaken by the appropriate actors, a tiered system of public disclosure has been proposed, where layman's summaries giving prescribed information are released to the public, while sensitive and detailed technical information may be withheld or redacted until a permitted request for further information is made, or the regulator requests it.[57]

The entity should publish the purpose, scope, intended use of the ADS. The entity should also disclose and explain the approach adopted; stakeholders affected or likely to be affected; impacts mitigated, and those addressed; and basic details about the AIA process itself. The step must be aimed at identifying pre-emptive actions.

Policy Briefing

Internal vs external audits

Internal auditing is conducted by individuals internal to the organisation, who have full access to the models, training data and key employees.[58] Insiders may sometimes be the best candidates for early detection of problems in the development of ADS.[59] As established norms in the algorithmic auditing world are sparse, there is high variance in the depth and quality of internal auditing. Raji proposes an end-to-end internal auditing framework, intended for integration with product life cycles, that bears much resemblance to the public agency model and the hiring Equality Impact Assessment (EQIA).[60] Raji's framework stresses the inclusion of internal knowledge and documentation, such as the principles, values and ethical objectives embodied in the development process.

Article 43 of the proposed EU AI Act mandates internally conducted ex ante risk assessments by providers of high-risk systems only, as part of the conformity assessment procedure. Although elements of the public agency model have been incorporated, including disclosing the purpose of the system, methods and key design choices, conformity assessments have been roundly criticised for having inconsistent thresholds, overreliance on internal auditing, and paying inadequate attention to those affected by AI systems.[61] Article 61 mandates post-market 'monitoring plans.' As a general heuristic, internal auditing incurs less external credibility but may grant better access, while external auditing grants less access but more robust external accountability.[62]

External auditing involves a third party auditor who is able to access parts of the ADS such as the outputs and the backend.[63] This can be done by the regulator or by a contracted third party. Regulators must be permitted to undertake full investigations with access to source codes, training datasets, methodologies, processes and techniques, as highlighted in IFOW's Machine Learning Case Studies.[64] The regulators must also be able to undertake the range of technical and non-technical audits and ongoing monitoring themselves, including code, scraping and API audits, as highlighted by the Ada Institute and Reset.[65]

Legislative norms and requirements for third party auditors could establish a healthy ecosystem for external AIAs and audits[66] by specifying bottom lines and level of access to encourage high standards, formalise external access privileges and prevent adversarial behaviour on the part of organisations. This ecosystem must enable such assessment by non-profit independent academic labs and civil society organisations, as well as the private sector. The specification dilemma has been coined to illustrate the invocation of model parameters to counter the findings of external auditors, who may not be granted full access to internal design specifications. However, authorised bodies could be granted full access on terms, which may enhance both accountability and reach to support innovation.[67]

In practice, key human decision making stages[68] will sit across varied organisations. In this context different parts of an impact assessment will be conducted 'internally' or 'externally'. Transparency and cooperation obligations across the supply chain will be required so that requests for relevant information or to make reasonable adjustments can be implemented.[69]

There are advantages of internal audits, but legislation must enable high quality, independent, external audits. Given current limitations of both, new requirements for auditing as part of the AIA must be combined with additional investigative powers to ensure regulators have full access to all relevant information and may undertake technical and non-technical interrogation where appropriate.

Policy Briefing

Quantitative vs qualitative impact assessment

Types of audit, metrics and methodological approach abound.[70] Some variables, and interpretations of fairness, are better suited to quantification but methodological variables cannot and should not delimit the scope of impact assessment. Broadly, audits can be differentiated by their quantitative or qualitative leanings.

Quantitative ex ante audits are typically conducted to test the accuracy of models and frame subjective questions of 'bias', or comparative outcomes between groups, as statistical metrics.[71] Quantitative audits are typically performed following system deployment (if at all) but can be delivered ex ante through using simulated testing environments or examining prior use cases.[72] IFOW's AI and Hiring paper and the CDEI Bias Review identify the limitations and inconsistencies of technical auditing in isolation, as well as the need for consistent, open and shared approaches to addressing bias, fairness and inequality. Particular attention must be given to auditing for equality, based on a comprehensive understanding of how this is distinct from 'bias' in relation to data processing.[73]

Qualitative ex ante audits involve the collection of qualitative data from developers, users and affected communities to draw a more holistic picture of impact. Information that could be obtained from organisations and development teams include justifications, assumptions, aims, wider organisation processes and culture, in the form of statements, declarations, interviews and ethnographic methodologies.[74] The use of qualitative data is intended to distil quantitative information such as model parameters and data processing in a descriptive, plain-language format, as well as uncover the reasoning and assumptions behind developer decision-making.

Several scholars propose an approach of rigorously examined alternatives, such as alternative variables, methods of assessment and mitigation, including the alternative of taking no action.[75] The sketching of alternatives is intended to allow external reviewers a greater understanding of the available options at hand to contextualise developer decision-making.

Both quantitative and qualitative methods must be part of an effective AIA. Auditing for equality needs particular mention and attention in new regulation.

Measurements of impact and identifying harms

A variety of processes and impacts across the lifecycle of the ADS can be evaluated to produce risk assessments. Six primary types of transparency have been identified, which could constitute different parts or types of risk assessment.[76] Disclosures should therefore include explanations of decision rationale; responsibility and chains of development, management and implementation; data and how it has been used for decision-making; fairness, including procedures to mitigate bias and assessment of whether individuals have received equal treatment; safety and performance, primarily in reference to technical indices such as accuracy and robustness; and impact, including monitoring procedures and impacts on individuals or society.

The definition and analysis of impact and harm on affected communities does not have strong norms and are likely to be subject to normative contestation. Ultimately, defining harm is a normative exercise with little regulatory precedent.[77] There is a lack of a universal definition of fairness, for example, and there are trade-offs and conflicts between differing statistical interpretations of fairness, as well as between desirable outcomes such as fairness and accuracy.[78] This suggests normative uncertainty around 'fairness', pointing to the need for both qualitative auditing and participatory mechanisms to enable normative

Policy Briefing

evaluation and analysis of risk and impact. In the light of this, the EPRS recommends focusing on allocative and representational harms.[79]

Identifying potential harms through international human rights standards has been proposed[80] and there is growing interest in mandating human rights impact assessment as part of AIAs.[81] Our review suggests that human rights impact assessments may offer a familiar lens through which to start identifying impacts, noting that many common harms derive from a rights-based framework. The Turing Institute has made detailed proposals to inform best practice, which would support this approach.[82] However, it is not exhaustive and likely to miss collective and some intersectional harms. This means that assessment should start with evaluation of impacts on human rights, equality of opportunity and disparity of outcome, and go on to consider other impacts that may not be caught by a human-rights based approach, such as security of information when data processing is not involved, and socio-economic and place-based disadvantage.

It can be challenging to identify the origins and source of adverse impacts, especially when ADS tend to have complex supply chains.[83] Where proxies are used to classify individuals, mixed expertise is required to identify how such proxies might be enacting bias. This is why qualitative ex ante auditing is important for uncovering intentions, assumptions and purposes of a system.

IFOW research suggests that the harms that are most relevant in a work context are those deriving from legal and socio-economic rights and principles established under national and international law, applied and synthesised in the Charter of Good Work.[84] These are protected in the International Covenant on Economic, Social and Cultural Rights, among other legal instruments.[85] A sharp focus on work is also consistent with a recent emphasis on the social and economic implications of algorithmic systems in AI Ethics. We note that some measures for wider dimensions of work quality have already been developed, for example wellbeing measures in relation to auditing autonomous systems.[86] The Charter could therefore be used as a checklist of impacts on work, with metrics developed to combine these approaches (rights-based, social and economic interests and wellbeing measures). Standardised metrics, methods and techniques can be developed over time.[87]

Equality as a case study for risk and impact analysis

AI at Work offers huge potential to understand and correct historic patterns of inequality, if this is prioritised as an objective within the AIA framework. However, AI models are trained on data that reflects past patterns of behaviour and allocations of resource, and can therefore project these patterns of inequality into the future at an unprecedented scale and pace, unless this is consciously corrected for.[88]

While research on the best approach and methods to counter the reproduction of inequalities is in progress, current work broadly supports a three step process to identifying and mitigating relevant risks to equality of opportunity and outcome on individuals and groups.[89] These steps should be integrated into the AIA process. The literature suggests that there are several, viable approaches to assessment for each step, which point to regulation setting a broad goal and matters which must be considered. Precise definitions, remit or methodology should be disclosed but not mandated, and detailed guidance must follow.

Firstly, risks, impacts and anticipated impacts on groups with shared protected characteristics would be examined. This reflects the current approach in rights-based frameworks, in particular the Equality Act, as highlighted by Robin Allen QC and Dee Masters.[90]

Secondly, the AIA would require due regard to the desirability of reducing inequalities of outcome resulting from socio-economic and place-based disadvantage, given well-established limitations of assessing equality only by reference to single, protected characteristics, as recognised in the CDEI's Bias Review.[91] The model in s1 Equality Act is instructive here.

Lastly, a wider review would be made in response to evidence emerging from the participatory forums and unrestricted by specific, anticipated outcomes. Future-proof regulation must allow for assessment and response to new and unexpected impacts.

New AI capabilities require that particular attention is given to assessing equality of opportunity and outcome, especially with regard to socio-economic and place-based disadvantage. Existing legal models, combined with requirements for participatory mechanisms, offer alternatives to requiring proof of causation.

Stage 3

Taking appropriate action in response to the ex ante analysis

Both the literature review and case studies point to the importance of establishing a process for AIAs aimed at enabling appropriate mitigation. Existing tools have not been built for this purpose, which points to the need for clear regulation, with better tools and guidance to support practical applications.[92] This should start with requirements for such a process, and the formulation of a plan. There is a sound case for this to allow access to external researchers, in both the technical and social science fields, to review the system once it is deployed.[93]

Public engagement may help identify priorities, routes and trade-offs, enabled through public comment periods and external researcher review.[94] The organisation could publish the AIA as a single document with one subsequent participation period, or divide the process into separate publication and participation periods, structured according to the public agency model components (definition, disclosure, self-assessment/external audit and meaningful access). In response to public or stakeholder comment, the organisation can make further reasonable adjustments to the ADS model or organisation policies to address concerns.

The public agency model invites release of the final version of the AIA, allowing for public challenge in the case of failure or inadequacy to mitigate issues raised in the public participation period. A due process challenge period should be implemented where the public, particularly affected communities, have a further opportunity to bring concerns to an oversight body, regulatory agency, or in the courts if none or insufficient action is taken.[95]

The literature points to the particular challenges of establishing causation for representational and allocative harms, in particular, where large numbers of people and entities are involved.[96] This challenge could be addressed in regulation by drawing on the legal model of contributory negligence, where there is a wide measure of judicial discretion to explore factors relevant to causation and apportion 'just and equitable' compensation in the circumstances of the case.[97] Relevant facts and material would have to be considered, subject to caveat of availability if reasonable steps are taken to obtain it, an approach which draws from the UK's model of judicial review.[98] The recent landmark case on environmental harms is also instructive, pointing to the importance of actual and imputed knowledge, and of published guidance, by the UN in the case of Shell.[99]

This points to the value for business, as well as stakeholders, in establishing a thorough process and transparent, effective mechanisms to enable consultation and wider evidence-gathering so that harms and other impacts are understood, identified and considered in practice, followed by context-specific adjustments. Reasonable adjustments should take into account the resources and capabilities of an organisation, alongside proximity to the harm and its severity, as modelled by the law requiring reasonable adjustments for disability.[100] Detailed guidance on matters to be taken into account and trade-offs should follow, once the principle is established. Overall, adjustments should advance application of the overarching principles.

Stage 4

Continuous evaluation to ensure assessment and appropriate action is ongoing

The final stage of the AIA is to ensure continuous evaluation. Various forms of impact will only be identifiable once a system is deployed in practice. For this reason, the recommendation of ongoing monitoring of algorithmic systems is a near-universal recommendation amongst scholars. Once new impacts are identified, new mitigations should be put in place. This approach recognises a wide range of potential impacts, which may be evaluative, representative widely allocated. Research on the 'gigification' of work suggests that these may only become apparent over time and can be profound, observed across all 10 dimensions of good work.[101]

Establishing a mechanism for ongoing evaluation will enable participatory and reflexive exercises to deepen understanding of impacts and potential mitigative measures that are responsive to the context at hand.[102] This will enable improved iterations of the assessment as practice, guidance and caselaw builds on conducting AIAs in the public interest.

IFOW case studies and the literature suggests that an AIA process should be dynamic and renewed in their entirety on a regular basis, such as every year or two years, and when there is a change of purpose or remit of the algorithmic system.[103] Machine Learning case studies suggest that specific triggers should include:

- Any 'further processing' of data;
- Any new third party interactions;
- Any change of use within the system, which employers should persistently track;
- Any addition of new datasets/sources.

“

As a practical response to this deficit of responsible foresight, I implore you to consider mandating regimes of impact assessment.

Dr David Leslie, Ethics Theme Lead at The Alan Turing Institute

Quote from the APPG on the Future of Work’s inquiry into the New Frontier.

Conclusion

As algorithmic systems are increasingly used by governments and businesses to make automated, data-driven decisions, legislation must not only keep pace with these changes but future-proof mechanisms for accountability and shape future directions in the public interest.

This proposal comes at a time when global agreement on AI Ethics marked by the UNESCO guidance this month provides a normative basis for codification of a principle-based framework in regulation grounded in new duties to undertake and act on AIAs. Based on insight from the world of work, our briefing has shown why and how AIAs could become an overarching framework for the assurance and accountability of algorithmic decision making and support systems. This approach would allow the UK Government to achieve the vision of the AI Strategy and lead in setting global standards for the development of responsible technology in the work place and beyond.

Annexes

Annex 1 Draft amendments on AIAs[104]

Statement of purpose

(1) Any organisation applying a relevant algorithmic above must make a statement of purpose including

(a) a summary description of the algorithmic system

(b) its capabilities, remit and proposed applications

(c) how to access further information

(2) The statement of purpose made under (1) should be available

(a) in public annual and regulatory reports.

(b) on request of any person or group of persons under s1

(3) An organisation applying a relevant algorithmic system must notify a person and any designated representatives of the statement of purpose where any decision has been made concerning:

(a) hiring and access to work

(b) pay or work allocation

(c) monitoring or evaluation of performance

(d) discipline or termination of work

Tiers of disclosure

(1) A party should record documentation relevant to development, procurement and/or application of an algorithmic system including

(a) the identification of third parties contracted

(b) the outcome proposed

(c) the programming, training methodologies and data used

(d) the techniques used to test and validate the system

(e) the variables and weighting of variables selected to predict, rank or classify the outcome

(f) any trade offs between different measures and given rationale

(g) any organisational policies and processes relevant to procurement and applications of the system

(2) person with a relevant interest, right or freedom may request a summary statement of the matters documented under a, b and g at any time after deployment of the system

(3) On receipt of a summary statement, a person or designated representative of a group of persons with a relevant interest, right or freedom may request a written explanation of how the summary statement under s2 applies to him within 3 months of receipt of the statement.

(4) Prescribed organisations may request copies of relevant documentation identified for the purpose of research, development or undertaking an independent algorithmic impact assessment.

(5) Nothing in this section detracts from any other rights and duties for record, inspection or disclosure.

Annex 1

Algorithmic Impact Assessments

(1) Prior to use or procurement of relevant algorithmic system an organisation is responsible for completing an Algorithmic Impact Assessment in a form prescribed by regulations made under this Act.

(2) The Algorithmic Impact Assessment must be reviewed and updated at regular intervals of every 2 years and/or when the functionality, or the scope, of the automated decision system changes

(3) The Algorithmic Impact Assessment must:

(i) identify the person and groups of persons sharing a relevant interest

(ii) analyse the risks and impacts of the system including impacts on

(a) equality of opportunity, disparity of outcome and human rights

(b) safety, privacy and security whether or not personal data is processed

(c) good work

(d) any other adverse impacts identified in the course of the assessment process and/or subsequent monitoring

(iii) include a technical audit

(iv) provide a statement of the period for wider consultation and process for stakeholder participation

(v) identify potential adjustments or other steps that could be taken in response to the evaluation

(4) Any person with a relevant interest, right or freedom may request a summary Algorithmic Impact Assessment at any time after deployment of the system

(5) The organisation undergoing the assessment should take into account any guidance and/or tools for Algorithmic Impact Assessment published by the CDEI, DCRF or regulators.

(6) Nothing in this part detracts from existing regulator or governance obligations for impact assessment although the algorithmic impact assessment may operate to discharge these obligations.

(7) In this section 'good work' means work with provides or promotes:

(a) fair access

(b) fair pay

(c) fair terms and conditions

(d) equality

(e) dignity

(f) autonomy

(g) physical and mental wellbeing

(h) access to institutions and people who can represent workers' interests

(i) participation to determine and improve working conditions

(j) access to facilities for career guidance and training

Annex 1

Reasonable steps

(1) An organisation with responsibility for a relevant algorithmic system must

(a) only consider relevant information recorded *[under disclosure provisions]*

(b) consider technical and non-technical mitigation or other actions *[under algorithmic assessment procedure]*

(c) take reasonable and proportionate steps to mitigate risks, address harms and promote benefits in the circumstances of the case including the size, resources, capabilities, severity of harm and proximity of the organisation to any adverse impact identified

(2) In determining reasonable steps under s 1, an organisation with responsibility for an algorithmic system person must have particular regard to

(a) impacts on vulnerable persons or groups that may be disproportionately affected

(b) the desirability of reducing inequalities of outcome which result from socio-economic and/or place-based disadvantage

(c) the desirability of promoting the international reputation of the United Kingdom and businesses in the UK for responsible innovation and adherence to the United Nations Sustainable Development Goals and the UNESCO Agreement on Artificial Intelligence

Duty to co-operate

(1) An organisation required to record documentation must in contracting with any party to develop, procure of apply any part of the algorithmic system ('a contracted party') secure agreement by the contracted party to provide on request by the organisation

(a) the documentation required in order to produce an Algorithmic Impact Assessment *[see above]*

(b) the documentation required in order to take reasonable steps *[see above]*

(2) A party's duty to record documents is limited to documents which are or have been in his control

For this purpose a party has or has had a document in his control if –

(a) it is or was in the organisation's possession

(b) the organisation has or has had a right to possession of it or

(c) the organisation has or has had a right to request, inspect or take copies of it.

(3) Organisations may redact material that is commercially sensitive *and* irrelevant to the decision under consideration except under s 4 of this section

(4) The Digital Cooperation Regulation Forum or individual regulators may request all documentation and other information from any organisation or party contracted to undertake any part of development, procurement or application of a relevant algorithmic system.

Annex 1

Designing for accountability

(1) A relevant algorithmic system must be designed to take into account

(a) interpretability of decisions made or supported

(b) equality of opportunity, disparity of outcome and human rights

(c) safety, privacy and security of personal information

(d) good work

(2) In this section 'good work' means work with provides or promotes::

(a) fair access

(b) fair pay

(c) fair terms and conditions

(d) equality

(e) dignity

(f) autonomy

(g) physical and mental wellbeing

(h) access to institutions and people who can represent workers' interests

(i) participation to determine and improve working conditions

(j) access to facilities for career guidance and training

Relevant algorithmic systems

(1) This Part applies to any algorithmic system developed, procured or applied by an organisation six months after the date on which this Act is passed where there is a reasonable prospect of engaging

(a) a person's interests, rights and/or freedoms

(b) a group of persons sharing an interest under (a)

(2) Relevant algorithmic systems include any system, tool or model that is developed, procured or applied to make or inform a decision relating to

(a) access, terms or conditions of work

(b) tasks undertaken at work

(c) opportunities for learning, promotion or other benefits

is a relevant algorithmic system under this section.

(3) Any algorithmic system involving supervised, unsupervised or reinforcement machine learning procured or deployed at work

is a relevant system under this section.

(4) Nothing in this Part detracts from other duties, rights, freedoms or interests of a person in any other primary or secondary legislation relating personal data, employment or social protection.

Annex 2 The Good Work Charter

1 **Access**
Everyone should have access to good work

2 **Fair pay**
Everyone should be fairly paid

3 **Fair conditions**
Everyone should work on fair conditions set out on fair terms

4 **Equality**
Everyone should be treated equally and without discrimination

5 **Dignity**
Work should promote dignity

6 **Autonomy**
Work should promote autonomy

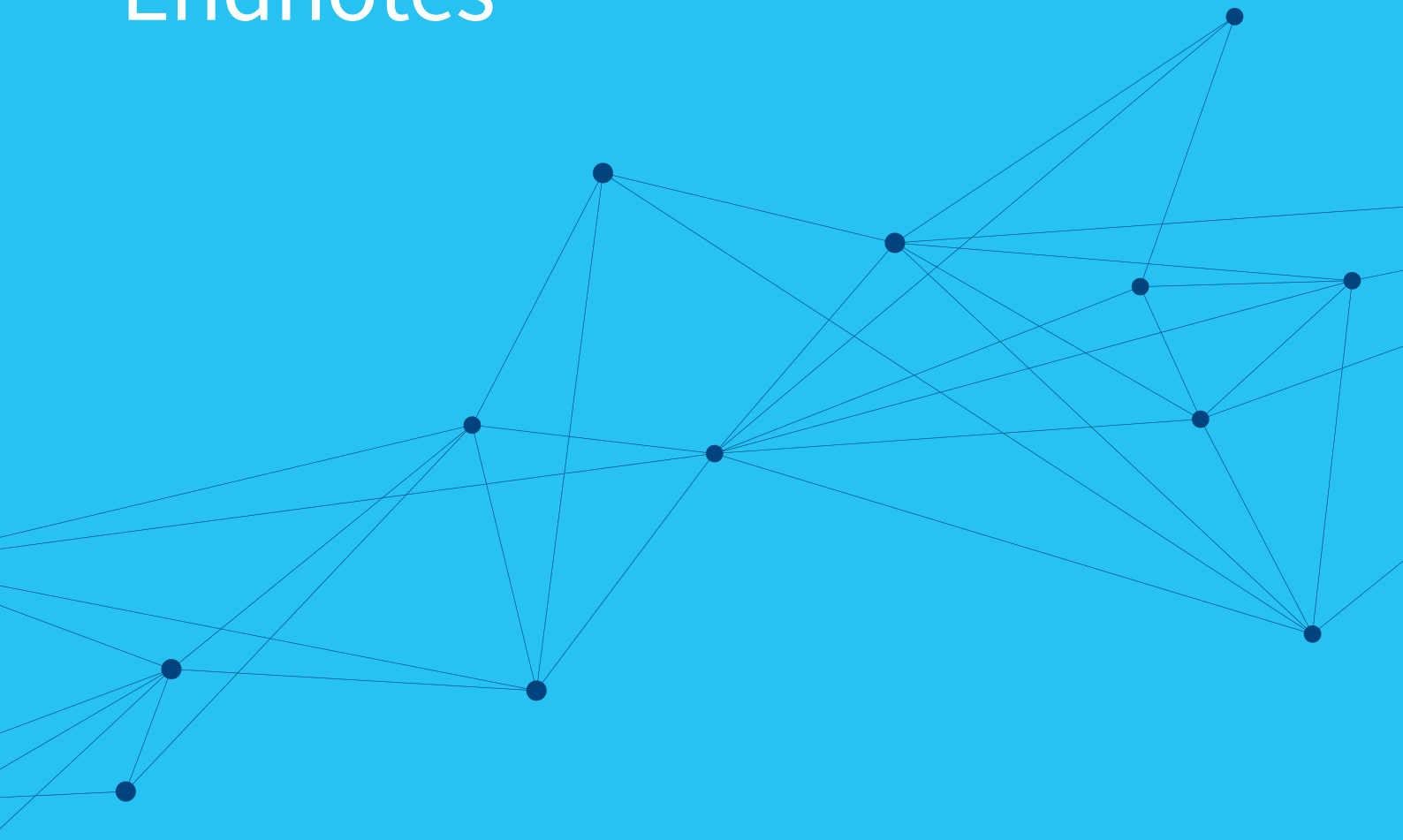
7 **Wellbeing**
Work should promote physical and mental wellbeing

8 **Support**
Everyone should have access to institutions and people who can represent their interests

9 **Participation**
Everyone should be able to take part in determining and improving working conditions

10 **Learning**
Everyone should have access to lifelong learning and career guidance

Endnotes



Endnotes

1 Department for Digital, Culture, Media and Sport. AI Strategy, 2021.

2 CDEI 'Engaging with the public about algorithmic transparency in the public sector' 2021.

3 Department for Digital, Culture Media and Sport 'Data: A New Direction' consultation, 2021.

4 Gilbert, Abigail and Anna Thomas 'The Amazonian Era: The Gigification of Work' Institute for the Future of Work, 2021.

5 Graham, Logan, Abigail Gilbert, Joshua Simons, and Anna Thomas. 'Artificial Intelligence in Hiring'. Institute for the Future of Work, 2020.

6 Department for Digital, Culture, Media and Sport. Draft Online Safety Bill, 2021.

7 Department for Digital, Culture, Media and Sport. Draft Online Safety Bill, 2021.

8 Algorithmic Accountability for the Public Sector: Learning from the First Wave of Policy Implementation.

9 Centre for Data Ethics and Innovation. 'Review into Bias in Algorithmic Decision-Making', 2020.

10 All-Party Parliamentary Group on the Future of Work. 'Report into The New Frontier: Artificial Intelligence at Work', 2021.

11 Graham, Logan, Abigail Gilbert, Joshua Simons, and Anna Thomas. 'Artificial Intelligence in Hiring'. Institute for the Future of Work, 2020.

12 Department for Digital, Culture Media and Sport 'Data: A New Direction' consultation, 2021.
Businesses are increasingly concerned about algorithms causing major fiscal or repetitional damage. Koshiyama, Adriano, Emre Kazim, Philip Treleaven, Pete Rai, Lukasz Szpruch, Giles Pavey, Ghazi Ahamat et al. "Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms." (2021). The need for effective 'assurance' has been flagged by the CDEI, https://cdei.blog.gov.uk/2021/04/15/the-need-for-effective-ai-assurance/ with impact assessment a recognised part of this.

13 Gilbert, Abigail and Anna Thomas 'The Amazonian Era: The Gigification of Work' Institute for the Future of Work, 2021.
Recruitment presents a particularly challenging use case for auditing impacts, see: Kazim, Emre, Adriano Soares Koshiyama, Airlie Hilliard, and Roseline Polle. "Systematizing Audit in Algorithmic Recruitment." *Journal of Intelligence* 9, no. 3 (2021): 46.

14 Moss, Emanuel, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 'Assembling Accountability: Algorithmic Impact Assessment for the Public Interest'. Available at SSRN 3877437, 2021.

15 Ada Lovelace Institute, AI Now Institute, and Open Government Partnership. 'Algorithmic Accountability for the Public Sector', 2021. https://www.opengovpartnership.org/documents/ algorithmic-accountability-public-sector/.

16 Ad Hoc Committee on Artificial Intelligence (CAHAI). 'Feasibility Study'. Strasbourg: Council of Europe, 2020.

17 Selbst, Andrew D. 'An Institutional View Of Algorithmic Impact Assessments', 2021.

18 Department for Digital, Culture, Media and Sport. 'Data: A New Direction', 2021.

19 Kaminski, Margot E., and Gianclaudio Malgieri. 'Algorithmic Impact Assessments under the GDPR: Producing Multi-Layered Explanations'. SSRN Electronic Journal, 2019. https://doi.org/10.2139/ssrn.3456224.
However, the DPIA model as a form of impact assessment also only considers a limited range of harms, see more in: Kazim, E., Koshiyama, A. The interrelation between data and AI ethics in the context of impact assessments. AI Ethics 1, 219–225 (2021). https://doi.org/10.1007/s43681-020-00029-w.

20 Information Commissioners Office 'Consultation on the ICO's AI and data protection risk toolkit' 2021 https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/consultation-on-the-ico-s-ai-and-data-protection-risk-toolkit/.

21 Binns, Reuben, Abigail Gilbert, Anne-Marie Imafidon, Tim Johnston, David Leslie, Joshua Simons, Helen Mountfield, and Anna Thomas. 'Mind the Gap: How to Fill the Equality and AI Accountability Gap in an Automated World'. Institute for the Future of Work, 2020.

22 Prospect and Institute for the Future of Work 'Data Protection Impact Assessments: A Union Guide' 2020.

23 Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. 'Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability'. AI Now Institute, 2018, 1–22.

24 Binns, Reuben, Abigail Gilbert, Anne-Marie Imafidon, Tim Johnston, David Leslie, Joshua Simons, Helen Mountfield, and Anna Thomas. 'Mind the Gap: How to Fill the Equality and AI Accountability Gap in an Automated World'. Institute for the Future of Work, 2020.

25 Information Commissioners Office 'Call for Views on Employment Practices', 2021.

Endnotes

26 Mökander, Jakob, Maria Axente, Federico Casolari, and Luciano Floridi. 'Conformity Assessments and Post-Market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation', 2021, 27; Moss, Emanuel, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 'Assembling Accountability: Algorithmic Impact Assessment for the Public Interest'. Available at SSRN 3877437, 2021.

27 Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. 'Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability'. AI Now Institute, 2018, 1–22; European Parliamentary Research Service. A Governance Framework for Algorithmic Accountability and Transparency. LU: Publications Office, 2019. https://data.europa.eu/doi/10.2861/59990; Kaminski, Margot E., and Gianclaudio Malgieri. 'Algorithmic Impact Assessments under the GDPR: Producing Multi-Layered Explanations'. SSRN Electronic Journal, 2019. https://doi.org/10.2139/ssrn.3456224; Selbst, Andrew D. 'An Institutional View Of Algorithmic Impact Assessments', 2021.

28 Simons, Josh, and Anna Thomas. 'Machine Learning Case Studies'. Institute for the Future of Work, 2020.

29 Moss, Emanuel, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 'Assembling Accountability: Algorithmic Impact Assessment for the Public Interest'. Available at SSRN 3877437, 2021. Field experience from UCL and Holistic AI suggest that Human Resource Impact Assessments are often, in practice, reduced to questions of compliance with Equality Legislation, which itself requires greater clarification to manage trade-offs.

30 Simons, Josh, and Anna Thomas. 'Machine Learning Case Studies'. Institute for the Future of Work, 2020.

31 United States Congress. Algorithmic Accountability Act of 2019, 2019.

32 Information Commissioner's Office. 'Examples of Processing "Likely to Result in High Risk"', n.d.

33 Treasury Board of Canada Secretariat, Policy on Service and Digital.

34 Ibid.

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

36 Binns, Reuben, Abigail Gilbert, Anne-Marie Imafidon, Tim Johnston, David Leslie, Joshua Simons, Helen Mountfield, and Anna Thomas. 'Mind the Gap: How to Fill the Equality and AI Accountability Gap in an Automated World'. Institute for the Future of Work, 2020.

37 Information Commissioner's Office. 'Examples of Processing "Likely to Result in High Risk"', n.d.

38 Moss, Emanuel, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 'Assembling Accountability: Algorithmic Impact Assessment for the Public Interest'. Available at SSRN 3877437, 2021.

39 Binns, Reuben, Abigail Gilbert, Anne-Marie Imafidon, Tim Johnston, David Leslie, Joshua Simons, Helen Mountfield, and Anna Thomas. 'Mind the Gap: How to Fill the Equality and AI Accountability Gap in an Automated World'. Institute for the Future of Work. Also, evidence provided by Dr David Leslie of the Turing Institute to the All Party Parliamentary Group on the Future of Work Inquiry into the use of AI at the Workplace.

40 Leslie, David. 'Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems: A Proposal Prepared for the Council of Europe's Ad Hoc Committee on Artificial Intelligence', n.d; All-Party Parliamentary Group on the Future of Work. 'Report into The New Frontier: Artificial Intelligence at Work', 2021.

41 Binns, Reuben, Abigail Gilbert, Anne-Marie Imafidon, Tim Johnston, David Leslie, Joshua Simons, Helen Mountfield, and Anna Thomas. 'Mind the Gap: How to Fill the Equality and AI Accountability Gap in an Automated World'. Institute for the Future of Work, 2020.

42 Rieke, Aaron, Miranda Bogen, and David Robinson. 'Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods'. Omidyar Network; Upturn, 2018.

43 Cobbe, Jennifer, Michelle Seng Ah Lee, and Jatinder Singh. 'Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems'. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 598–609. Virtual Event Canada: ACM, 2021. https://doi.org/10.1145/3442188.3445921.

44 We note other domains of impact assessments could have more established norms in this regard.

45 Leslie, David. 'Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector'. Available at SSRN 3403301, 2019.

46 Katyal, Sonia K. 'Private Accountability in an Age of Artificial Intelligence'. UCLA Law Review 66 (2019): 54–141.

47 Katyal, Sonia K. 'Private Accountability in an Age of Artificial Intelligence'. UCLA Law Review 66 (2019): 54–141.

Endnotes

48 Kaminski, Margot E., and Gianclaudio Malgieri. 'Algorithmic Impact Assessments under the GDPR: Producing Multi-Layered Explanations'. SSRN Electronic Journal, 2019. https://doi.org/10.2139/ssrn.3456224.

49 HUDERIAs are the choice of governance mechanism for the majority of stakeholders in a recent surevy: 81% of respondents selected "Human rights, rule of law and democracy impact assessments", followed by "audits and intersectional audits" (70%) and mechanisms of "certification and quality labelling" (51%). Council of Europe's public survey on how to define and regulate. https://ecnl.org/news/ai-policy-message-delivered-will-states-listen.

50 International Finance Corporation. 'Stakeholder Engagement: A Good Practice Handbook for Companies Doing Business in Emerging Markets'. World Bank Group, 2007.

51 See forthcoming guidance on stakeholder engagement for workplace audits of algorithmic systems by Institute for the Future of Work, 2022.

52 All Party Parliamentary Group on the Future of Work Inquiry into AI at the Workplace, Evidence from Dr David Leslie.

53 Simons, Josh, and Anna Thomas. 'Machine Learning Case Studies'. Institute for the Future of Work, 2020;
Gilbert, Abigail and Anna Thomas 'The Amazonian Era: The Gigification of Work'. Institute for the Future of Work, 2021.

54 Ibid and reference amendment.

55 Moss, Emanuel, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf.
'Assembling Accountability: Algorithmic Impact Assessment for the Public Interest'. Available at SSRN 3877437, 2021.

56 Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 'Datasheets for Datasets'. ArXiv:1803.09010 [Cs], 19 March 2020. http://arxiv.org/abs/1803.09010.

57 CDEI 'Engaging with the public about algorithmic transparency in the public sector' 2021.

58 Kazim, Emre, Adriano Soares Koshiyama, Airlie Hilliard, and Roseline Polle. 'Systematizing Audit in Algorithmic Recruitment'. Journal of Intelligence 9, no. 3 (17 September 2021): 46. https://doi.org/10.3390/jintelligence9030046.

59 Mökander, Jakob, Maria Axente, Federico Casolari, and Luciano Floridi. 'Conformity Assessments and Post-Market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation', 2021, 27.

60 Raji, Inioluwa Deborah, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes.
'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing', 2020, 12.

61 Mökander, Jakob, Maria Axente, Federico Casolari, and Luciano Floridi. 'Conformity Assessments and Post-Market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation', 2021, 27.

62 Sloane, Mona, Emanuel Moss, and Rumman Chowdhury. 'A Silicon Valley Love Triangle: Hiring Algorithms, Pseudo- Science, and the Quest for Auditability', 2021, 10.

63 Moss, Emanuel, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 'Assembling Accountability: Algorithmic Impact Assessment for the Public Interest'. Available at SSRN 3877437, 2021.

64 Binns, Reuben, Abigail Gilbert, Anne-Marie Imafidon, Tim Johnston, David Leslie, Joshua Simons, Helen Mountfield, and Anna Thomas. 'Mind the Gap: How to Fill the Equality and AI Accountability Gap in an Automated World'.
Institute for the Future of Work, 2020.

65 Ada Lovelace Institute and Reset 'Inspecting Algorithms in Social Media Platforms', 2020.

66 Ada Lovelace Institute and Reset 'Inspecting Algorithms in Social Media Platforms', 2020.
Third party auditors need to be competent. In fact, an ecosystem may be needed. See more in: Thelisson, Eva, Kirtan Padh, and L. Elisa Celis. "Regulatory mechanisms and algorithms towards trust in AI/ML." In *Proceedings of the IJCAI 2017 workshop on explainable artificial intelligence (XAI), Melbourne, Australia*. 2017; Barclay, Iain, Harrison Taylor, Alun Preece, Ian Taylor, Dinesh Verma, and Geeth de Mel. "A framework for fostering transparency in shared artificial intelligence models by increasing visibility of contributions." *Concurrency and Computation: Practice and Experience* 33, no. 19 (2021): e6129.

67 Moss, Emanuel, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 'Assembling Accountability: Algorithmic Impact Assessment for the Public Interest'. Available at SSRN 3877437, 2021.

68 Binns, Reuben, Abigail Gilbert, Anne-Marie Imafidon, Tim Johnston, David Leslie, Joshua Simons, Helen Mountfield, and Anna Thomas. 'Mind the Gap: How to Fill the Equality and AI Accountability Gap in an Automated World'.
Institute for the Future of Work, 2020.

69 See proposed amendment in the Annex; informed by Binns, Reuben, Abigail Gilbert, Anne-Marie Imafidon, Tim Johnston, David Leslie, Joshua Simons, Helen Mountfield, and Anna Thomas.
'Mind the Gap: How to Fill the Equality and AI Accountability Gap in an Automated World'. Institute for the Future of Work, 2020.

Endnotes

70 Ada Lovelace Institute, AI Now Institute and Open Government Partnership. (2021). Algorithmic Accountability for the Public Sector. Available at: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/.

71 Graham, Logan, Abigail Gilbert, Joshua Simons, and Anna Thomas. 'Artificial Intelligence in Hiring'. Institute for the Future of Work, 2020; Kazim, Emre, Adriano Soares Koshiyama, Airlie Hilliard, and Roseline Polle. 'Systematizing Audit in Algorithmic Recruitment'. Journal of Intelligence 9, no. 3 (17 September 2021): 46. https://doi.org/10.3390/jintelligence9030046.

72 Moss, Emanuel, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 'Assembling Accountability: Algorithmic Impact Assessment for the Public Interest'. Available at SSRN 3877437, 2021.

73 Binns, Reuben, Abigail Gilbert, Anne-Marie Imafidon, Tim Johnston, David Leslie, Joshua Simons, Helen Mountfield, and Anna Thomas. 'Mind the Gap: How to Fill the Equality and AI Accountability Gap in an Automated World'. Institute for the Future of Work, 2020.

74 Ada Lovelace Institute. 'Beyond Face Value: Public Attitudes to Facial Recognition Technology', 2019; Sloane, Mona, Emanuel Moss, and Rumman Chowdhury. 'A Silicon Valley Love Triangle: Hiring Algorithms, Pseudo- Science, and the Quest for Auditability', 2021, 10.

75 Katyal, Sonia K. 'Private Accountability in an Age of Artificial Intelligence'. UCLA Law Review 66 (2019): 54–141; Kaminski, Margot E., and Gianclaudio Malgieri. 'Algorithmic Impact Assessments under the GDPR: Producing Multi-Layered Explanations'. SSRN Electronic Journal, 2019. https://doi.org/10.2139/ssrn.3456224.

76 Leslie, David and Briggs, Morgan. 'Explaining Decisions Made with AI: A Workbook (Use Case 1: AI-Assisted Recruitment Tool)'. Zenodo, 20 March 2021. https://doi.org/10.5281/ZENODO.4624711.

77 Moss, Emanuel, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 'Assembling Accountability: Algorithmic Impact Assessment for the Public Interest'. Available at SSRN 3877437, 2021.

78 Kazim, Emre, Adriano Soares Koshiyama, Airlie Hilliard, and Roseline Polle. 'Systematizing Audit in Algorithmic Recruitment'. Journal of Intelligence 9, no. 3 (17 September 2021): 46. https://doi.org/10.3390/jintelligence9030046.

79 European Parliamentary Research Service. A Governance Framework for Algorithmic Accountability and Transparency. LU: Publications Office, 2019. https://data.europa.eu/doi/10.2861/59990. Allocative harms refer to the unequal distribution of resources across groups, while representational harms refers to the unequal distribution of groups in positions of power, influence and political or cultural representation.

80 Kaminski, Margot E., and Gianclaudio Malgieri. 'Algorithmic Impact Assessments under the GDPR: Producing Multi-Layered Explanations'. SSRN Electronic Journal, 2019. https://doi.org/10.2139/ssrn.3456224; Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. 'Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability'. AI Now Institute, 2018, 1–22.

81 Data & Society. 'Recommendations for Assessing AI Impacts to Human Rights, Democracy, and the Rule of Law', 2021.

82 Alan Turing Institute 'Human Rights, Democracy and the Rule of Law Assurance Framework for AI Systems: A Proposal Prepared for the Council of Europe's Ad hoc Committee on Artificial Intelligence' Forthcoming.

83 Hutchinson, Ben, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. "Towards accountability for machine learning datasets: Practices from software engineering and infrastructure." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 560–575. 2021.

84 Institute for the Future of Work, 'Good Work Charter' (2019). Available at: https://www.ifow.org/publications/the-ifow-good-work-charter.

85 In addition to right-based measures, some auditing impact measurements involve metrics relevant to workers but that are not rights-based, such as the IEEE 7010 standard which is a measure of wellbeing.

86 Schiff, Daniel, Aladdin Ayesh, Laura Musikanski, and John C. Havens. "IEEE 7010: A new standard for assessing the well-being implications of artificial intelligence." In *2020 IEEE international conference on systems, man, and cybernetics (SMC)*, pp. 2746–2753. IEEE, 2020.

87 Alan Turing Institute 'Human Rights, Democracy and the Rule of Law Assurance Framework for AI Systems: A Proposal Prepared for the Council of Europe's Ad hoc Committee on Artificial Intelligence' Forthcoming.

88 Eidelson, Benjamin. "Patterned Inequality, Compounding Injustice, and Algorithmic Prediction." *The American Journal of Law and Equality (Forthcoming)* (2021); Hellman, Deborah. "Personal Responsibility in an Unjust World: A Reply to Eidelson." *The American Journal of Law and Equality (forthcoming), Virginia Public Law and Legal Theory Research Paper* 2021–28 (2021). Such assessments should also question whether it is ultimately just and fair to automate certain processes, rather than whether or not automation is fair. This draws an important distinction between an engineering problem and a societal problem. Kazim, Emre, Jeremy Barnett, and Adriano Koshiyama. "Automation and Fairness: Assessing the Automation of Fairness in Cases of Reasonable Pluralism and Considering the Blackbox of Human Judgment." Available at SSRN 3698404 (2020).

Endnotes

89 In line with the mission of the Equality and Human Rights Commission.

90 Allen, Robin and Dee Masters. 'Technology Managing People – the legal implications: A report for the Trades Union Congress by the AI Law Consultancy' AI Law Hub, TUC 2021.

91 Centre for Data Ethics and Innovation, 'Review into Bias in Algorithmic Decision Making' 2020.
We further note issues in revealing intersectional bias and discrimination in evaluating either equality of opportunity or equality of outcome (when several demographic identifiers/protected charateristics are at play), when blunt statistical tests and classifications are deployed. This can lead organisations to have a mistaken sense of parity. Roseline Polle and Emre Kazim working via HolisticAI and UCL have had success in revealing intersectional discrimination in ML audits via application of Chi-Squared tests. This could provide technical grounds for a more nuanced reading of the Equality Act and permit more granular forms of mitigation.

92 Graham, Logan, Abigail Gilbert, Joshua Simons, and Anna Thomas. 'Artificial Intelligence in Hiring'.
Institute for the Future of Work, 2020.

93 Ada Lovelace Institute and Data Kind UK 'Examining the Black Box', 2020.
https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

94 Kaminski, Margot E., and Gianclaudio Malgieri. 'Algorithmic Impact Assessments under the GDPR: Producing Multi-Layered Explanations'. SSRN Electronic Journal, 2019. https://doi.org/10.2139/ssrn.3456224.

95 See also EqIA model in Graham, Logan, Abigail Gilbert, Joshua Simons, and Anna Thomas. 'Artificial Intelligence in Hiring'.
Institute for the Future of Work, 2020.

96 Suresh, Harini, and John Guttag. "A framework for understanding sources of harm throughout the machine learning life cycle."
In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–9. 2021.

97 Reed, Chris, Elizabeth Kennedy, and Sara Silva.
"Responsibility, autonomy and accountability: legal liability for machine learning."
Queen Mary School of Law Legal Studies Research Paper 243 (2016).

98 Cobbe, Jennifer.
"Administrative law and the machines of government: judicial review of automated public-sector decision-making."
Legal Studies 39, no. 4 (2019): 636–655.

99 Judgement by the Hauge District Court, Milieudefensie et al. v. Royal Dutch Shell plc.
https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2021:5339.

100 Equality Act 2010, c15 Part 2 Chapter 1 Section 6.
https://www.legislation.gov.uk/ukpga/2010/15/contents.

101 A Gilbert, Abigail and Anna Thomas 'The Amazonian Era: The Gigification of Work' Institute for the Future of Work, 2021.

102 Dobbe, Roel, Thomas Krendl Gilbert, and Yonatan Mintz.
"Hard Choices in Artificial Intelligence." arXiv preprint arXiv:2106.11022 (2021);
Lee, Min Kyung, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See et al.
"WeBuildAI: Participatory framework for algorithmic governance."
Proceedings of the ACM on Human-Computer Interaction 3, no. CSCW (2019): 1–35.

103 Kaminski, Margot E., and Gianclaudio Malgieri. 'Algorithmic Impact Assessments under the GDPR: Producing Multi-Layered Explanations'. SSRN Electronic Journal, 2019. https://doi.org/10.2139/ssrn.3456224;
European Parliamentary Research Service. A Governance Framework for Algorithmic Accountability and Transparency.
LU: Publications Office, 2019. https://data.europa.eu/doi/10.2861/59990.

104 Draft introduction for review: New clauses to require the establishment of a duty to undertake assessment of algorithmic impact assessments where organisations develop, procure or commence application of an algorithmic decision-making system; to disclose relevant information concerning the intended use of algorithmic decision-making system before procuring or commencing application of their use to persons likely to be affected by its use; to undertake consultation of persons likely to be affected before procuring or commencing application of an algorithmic decision-making system; to take reasonable steps to mitigate risks identified by such assessment before procuring or commencing application of an algorithmic decision-making system; to undertake regular biennial audits of the impact and effect of algorithmic decision making systems; to establish additional powers of investigation and remedies for the Digital Regulatory Forum; and for connected purposes.

Somerset House
The Exchange
London WC2R 1LA

www.ifow.org
@FutureWorkInst

Registered charity number 1180718