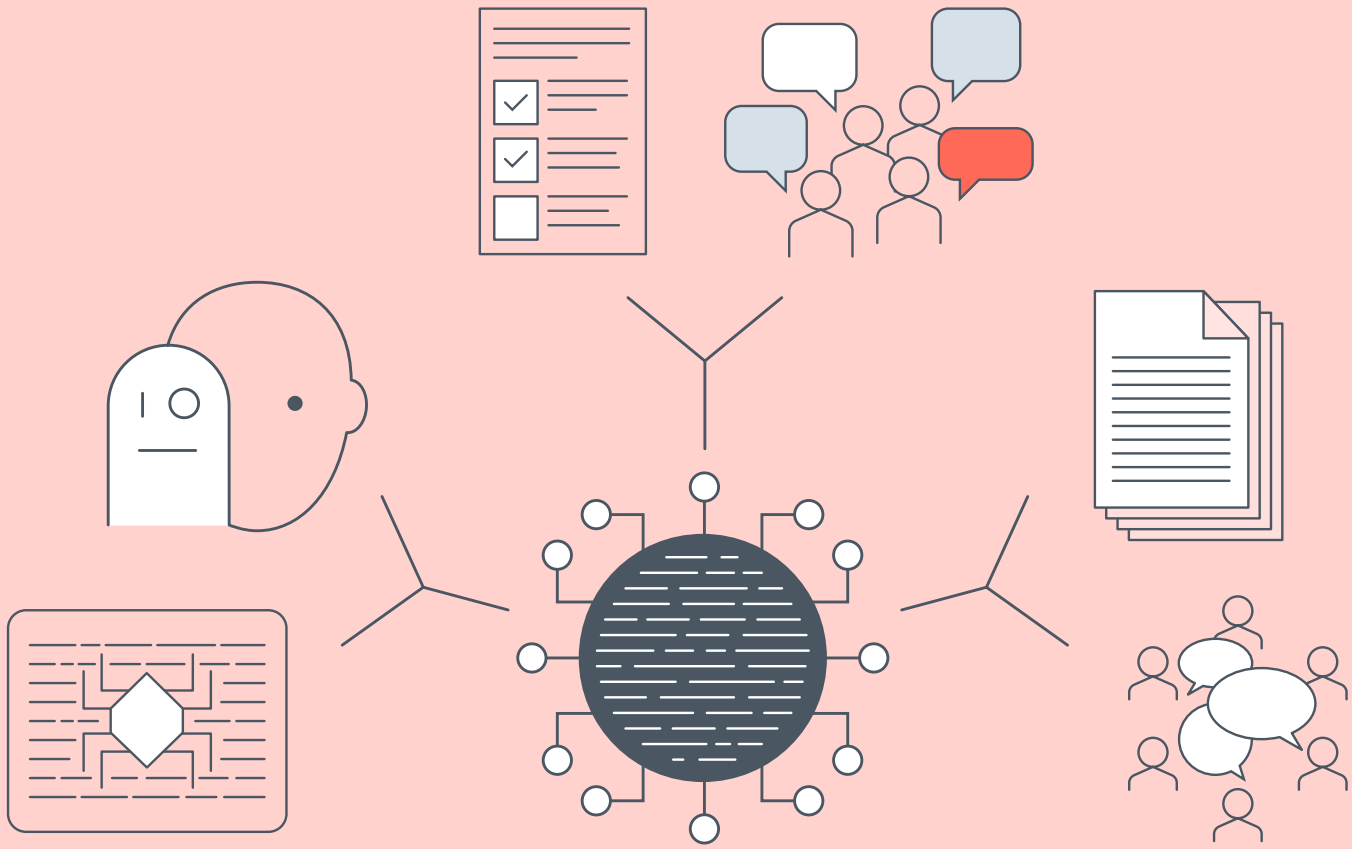
Ada Lovelace Institute

Law & policy

AI assurance?

Assesing and mitigating risks across the AI lifecyscle

July 2023



Contents

Executive summary

With the increasing use of AI systems in our everyday lives, it is essential to understand the risks they pose and take necessary steps to mitigate them. Because the risks of AI systems may become manifest at different stages of their deployment, and the specific kinds of risks that may emerge will depend on the contexts in which those systems are being built and deployed, assessing and mitigating risk is a challenging proposition.

Addressing that challenge requires identifying and deploying a range of methods across the lifecycle of an AI system's development and deployment.[1] By understanding these methods in more detail, policymakers and regulators can support their use in the UK's technology sector, and so reduce the risks that AI systems can pose to people and society.

In its March 2023 AI regulation white paper, the UK Government proposed creating a set of central Government functions to support the work of regulators. This included a cross-sectoral risk assessment function, intended to support regulators in their own risk assessments, to identify and prioritise new and emerging risks, and share risk enforcement best practices.

This central function has the potential to help coordinate and standardise the somewhat fragmented risk-assessment landscape identified in this paper and support the development of a cross-sectoral AI assessment ecosystem in the UK.

1 Ian Brown, *Allocating Accountability in AI Supply Chains* (Ada Lovelace Institute 2023) https://www.adalovelaceinstitute.org/resource/ai-supply-chains/

Key takeaways

1. **There is not a singular, standardised process for assessing the risks or impacts of AI systems** (or a common vocabulary), but there are commonly used components: policymakers, regulators and developers will need to consider how these are delivered and tailored.

2. **Risk and impact assessment methods typically involve five components:** risk identification, risk prioritisation, risk mitigation planning, risk monitoring and communicating risks. The main differences between components are in how they are achieved, the actors involved, the scope of impacts considered and the extent of accountability.

3. **Policymakers globally are incorporating risk and impact assessments in AI governance regimes and legislation**, with the EU, USA, Brazil and Canada mandating assessments for various AI systems. Regulators and policymakers face the challenge of ensuring risk consideration is conducted, acted on and monitored over time, highlighting the need for an ecosystem of assessment methods.

4. **Identifying and assessing risks alone does not ensure risks are avoided.** AI risk management will require an ecosystem of assessment, assurance and audit. This will include independent auditing, oversight bodies, ethics review committees, safety checklists, model cards, datasheets and transparency registers that collectively enable monitoring and mitigation of AI-related risks.

5. **Ensuring this AI assessment ecosystem is effective will require consensus on risk assessment standards**, supported by incentives for assessing societal risks and case studies showcasing risk-assessment methods in practice. Domain-specific guidance, skilled professionals and strong regulatory capacity can further enhance the ecosystem. Third-party assessors – including civil society, academia and commercial services – will be essential for developing and implementing assessment practices at scale.

Research questions

1. What are the broad areas of risks that AI systems can pose in different contexts (particularly from emerging AI technologies)?

2. How should regulators or policymakers respond to different kinds of risks?

3. What mechanisms and processes can be used to assess different kinds of risks, including the significance of their potential impact and their likelihood?

4. Whose responsibility (for example, developer, procurer, regulator) is it to conduct these assessments and evaluations?

5. What are methods for checking, monitoring and mitigating risks through the AI lifecycle?

6. What might be needed for an effective assessment ecosystem?

To answer these questions, this paper surveys approaches for assessing risks that AI systems pose for people and society – both on the ground within AI project teams, and in emerging legislation. The findings of this report are based on a desk-based review and synthesis of grey and academic literature on approaches to assessing AI risk, alongside analysis of draft regulations that contain requirements for anticipating risk or impacts of AI systems.

Key terms

Impact assessment: Impact assessments are evaluations of an AI system that use prompts, workshops, documents and discussions with the developers of an AI system and other stakeholders to explore how a particular AI system will affect people or society in positive or negative ways. These tend to occur in the early stages of a system's development before it is in use but may occur after a system has been deployed.

Risk assessment: Risk assessments are very similar to impact assessments but look specifically at the likelihood of harmful outcomes occurring from an AI system. These also tend to occur in the early stages of a system's development before it is in use but may occur after a system has been deployed.

Algorithmic audit: Algorithmic audits are a form of external scrutiny of an AI system, or the processes around it, which can be conducted as part of a risk or impact assessment. These can be technical audits of the inputs or outputs of a system; compliance audits of whether an AI development team has completed processes or regulatory requirements; regulatory inspections by regulators to monitor behaviour of an AI system over time; or sociotechnical audits that evaluate the ways in which a system is impacting wider societal processes and contexts in which it is operating. Audits usually occur after a system is in use, so can serve as accountability mechanisms to verify whether a system behaves as developers intend or claim.

Introduction

The last few years have seen a growing body of evidence of the risks AI systems can pose to people and society. In response, governments, industry organisations and civil society groups have developed a series of approaches for evaluating risks.

Each approach provides different methods for identifying potential risks of AI systems to particular groups, assessing the likelihood of those risks occurring and encouraging or suggesting interventions to mitigate them. However, there is currently little standardisation in approaches, and it can be challenging to navigate the range of approaches available.

This report surveys existing methods for assessing potential risks of AI systems – in literature and practice. It aims to support better understanding of how these methods can be used and the maturity of practice in specific areas, and to identify common or differentiating components of different methods. It also considers some wider mechanisms that can support monitoring and mitigation of those risks over time.

What we mean by AI risks

Risk can be thought of as a function of 1) the negative impact, or magnitude of harm, that would arise if the circumstance or event occurs and 2) the likelihood of occurrence.'[2] Negative impacts or outcomes can be experienced by individuals, groups, communities, organisations, society, the environment and the planet.

In all evaluations of risk, one of the most important questions to ask is 'a risk to who?'. For technology companies, risk may be understood in terms of business, reputational or financial risk. For policymakers, civil society and the public, risks to people and society may be front of mind.

2 Joint Task Force Transformation Initiative, 'Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy' (National Institute of Standards and Technology 2018) NIST SP 800-37r2 104 https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-37r2.pdf accessed 16 March 2023

Across different risk assessment approaches, the term 'impact' broadly refers to the potential positive or negative outcomes of a system, whereas 'risk' is focused on potential negative outcomes, and 'harm' refers to actual harmful outcomes that occur. However, in naming and discussion of methods these terms are sometimes used interchangeably.

Four ways to think about risks from AI systems

AI systems can pose a broad range of risks, but listing them all is a challenging task given the wide variety of contexts and sectors where AI systems can be deployed, and a lack of agreement over the definition of 'AI'. Nonetheless, we have identified four ways of thinking about risk and algorithmic harms in the literature:

1. **Risks of particular harms** such as representational harms, harms to equality, informational harms, physical or emotional harms, human rights infringement harms and societal harms. Some researchers differentiate between harms stemming from the design of an AI system and harms stemming from its use. For example, some differentiate between design flaws like faulty inputs or a failure to adequately test a system, and the risks in how an AI system is used (for example, to undermine civil and economic justice).[3] Others distinguish between risks caused by the ways a large language model (LLM) is trained (for example, lack of representation of non-English languages) and risks from how it is used (for example, to spread misinformation).[4]

2. **Risks associated with scenarios of AI systems** such as best- or worst-case scenarios, system failure, process failure or misuse.[5] Risks can occur not only when the AI system goes wrong, but also when the context around the system changes – and even when the system operates as intended. For example, researchers acknowledge that some social media algorithms are designed with

3 Rebecca Kelly Slaughter, Janice Kopec and Mohamad Batal, 'Algorithms and Economic Justice: A Taxonomy of Harms and a Path Forward for the Federal Trade Commission' (2021) *Yale Journal of Law & Technology* https://yjolt.org/sites/default/files/23_yale_j.l._tech._special_issue_1.pdf accessed 30 January 2023

4 Laura Weidinger and others, 'Taxonomy of Risks Posed by Language Models', *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2022) https://dl.acm.org/doi/10.1145/3531146.3533088

5 Ian Dafoe and Remco Zwetsloot, 'Thinking About Risks From AI: Accidents, Misuse and Structure' (*Lawfare*, 11 February 2019) https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure accessed 16 March 2023

the intention of extracting and monitoring user behaviour, which can pose an inherent risk to user privacy.[6] Others distinguish the risk of 'transfer context bias', in which an AI system designed for one context is inappropriately applied to another.[7]

3. **Risks associated with particular AI technologies**, where particular models or forms of AI systems have commonly associated risks. For instance, research has classified common risks of LLMs as discrimination,[8] hate speech and exclusion, information hazards, malicious uses, human-computer interaction harms and environmental and socioeconomic harms. Other approaches have classified the risks of deepfake technologies.[9]

4. **Risks associated with specific domains of application.** Risks from AI systems are dependent on the context in which they are deployed, such as healthcare or education. Some sectors are seeing domain-specific taxonomies of AI harms, like the economy,[10] or the environment.[11]

While these approaches to considering risk can offer regulators and policymakers a useful starting point for identifying potential harms an AI system may pose, identifying the risks raised by a particular AI system requires the use of risk assessment methods.

In the last few years, governments, industry organisations and civil society groups have produced a series of approaches for evaluating AI risks. Each approach provides different methods and techniques for identifying potential risks of AI systems to particular groups, the likelihood of those risks and steps to mitigate them. However, there is little standardisation of the methods used in each approach, and it can be challenging to navigate the range of approaches available.

6 Rebecca Kelly Slaughter, Janice Kopec and Mohamad Batal (n3)
7 Victor Galaz and others, 'Artificial Intelligence, Systemic Risks, and Sustainability' (2021) 67 *Technology in Society* 101741 https://www.sciencedirect.com/science/article/pii/S0160791X21002165
8 Laura Weidinger and others (n 4).
9 David Gray Widder and others, 'Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes', *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2022) https://dl.acm.org/doi/10.1145/3531146.3533779
10 Rebecca Kelly Slaughter, Janice Kopec and Mohamad Batal (n3)
11 Victor Galaz and others, (n 6)

Methods for assessing risks, outcomes and impacts

Risk assessment methods for AI systems are still emerging, with many under development. They vary in their scope of applications and in the specific prompts, questions and processes used by a risk assessor. They are currently rarely determined by consistent standards, and there is no consistent terminology for how to describe these methods, with some bodies describing them as 'frameworks' or 'toolkits'. Some of the activities described in risk and impact assessments also are also described in some methods as 'algorithm audits'.

The common theme within these methods is that they seek to anticipate harmful outcomes both before a system is in use and with the aim of monitoring or reassessing those risks as the system changes and develops. Risks from AI systems can manifest across their lifecycle and these systems can be dynamic – their behaviour can change with new inputs and data. When integrated into complex environments – like a hospital or a school environment – new risks can emerge over time. Therefore risk assessment approaches should also include ongoing monitoring.

These methods are beginning to appear as legal requirements in AI governance regimes like the Canadian Directive on Automated Decision-Making,[12] EU AI Act,[13] UK Online Safety Bill,[14] Brazil's draft AI legislation[15] and the proposed Algorithm Accountability Act in the USA.[16] These methods build on the use of risk and impact-assessment methods in

12 Treasury Board of Canada Secretariat, Directive on Automated Decision-Making 2019 https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592#cha6 accessed 21 February 2023

13 European Commission, Proposal for a Regulation Of The European Parliament and of the Council laying down harmonised Rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts 2021 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 accessed 21 March 2023

14 Michelle Donelan and Lord Parkinson of Whitley Bay, Online Safety Bill 2023 https://bills.parliament.uk/bills/3137 accessed 27 March 2023

15 Evangelos Sakiotis, Anna Oberschelp de Meneses and Nicholas Shepherd Quathem Kristof Van, 'Brazil's Senate Committee Publishes AI Report and Draft AI Law' [2023] *Inside Privacy* https://www.insideprivacy.com/emerging-technologies/brazils-senate-committee-publishes-ai-report-and-draft-ai-law/ accessed 28 March 2023;

16 Clarke, Text - H.R.6580 - 117th Congress (2021-2022): Algorithmic Accountability Act of 2022 2022

Risk assessment approaches should also include ongoing monitoring

data governance, such as data protection impact assessments (DPIAs) under EU and UK General Data Protection Regulation (GDPR). Many technology companies and public-sector bodies are also increasingly adopting these methods voluntarily as part of their governance process.

These methods also build on a history of practice in other fields, where they have been used as part of a governance process to assess the risks of potential policies and practices. For instance, impact assessments have a long history of use in finance, cybersecurity, data protection and environmental studies.[17] Similarly, risk-management approaches and assessments are common across business management, finance and assurance.[18]

When applied to AI, risk and impact-assessment methods aim to anticipate the impacts of AI systems and identify actions developers of AI systems can take so that risks can be mitigated and positive impacts can be best realised. Where risks are deemed to be too great, the assessments may indicate that a system is not appropriate for continued development, procurement or deployment.

Typically, these methods often lead to the creation of a final document that captures the results of the process. These methods share most or all of the following five components, though differ slightly in terms of: which actors are involved or responsible for each component; the scope of impacts considered; whether the assessment is voluntary or mandatory; and how the results of the assessment are communicated (see also table below):

1. **Risk identification:** Risk identification activities in both risk and impact assessments usually involve an exercise to compile answers to a set of prompts about the potential risks a system poses. This is often achieved through a workshop or discussion and captured in a document. Differences in technologies and contexts will determine who should be involved in the risk identification process (such as the development team, wider organisational stakeholders, external stakeholders or experts, user groups, or impacted communities).

17 Emanuel Moss and others, 'Governing with Algorithmic Impact Assessments: Six Observations' (24 April 2020) https://papers.ssrn.com/abstract=3584818 accessed 27 February 2023

18 Zenia Kotval and John Mullin, 'Fiscal Impact Analysis: Methods, Cases, and Intellectual Debate' (Lincoln Institute of Land Policy 2006) https://www.lincolninst.edu/sites/default/files/pubfiles/kotval-wp06zk2.pdf accessed 17 March 2023.

Risk identification activities can be driven by different prompts, for example: a particular set of risks (for example, risks to human or fundamental rights); the domain of application or technology used (for example, risks from medical AI or LLMs); or a scenario-led approach (for example, best-case outcome, worst-case outcome, system failure or system misuse) – or a combination of these.

2. **Risk prioritisation:** Risk prioritisation is often a combined weighting of risk likelihood, size, scope and affected parties (for example, a particular assessment may focus on risks to children). Prioritisation is a subjective activity – priority of risks depends on risks to who (which individuals or groups of people), and how those affected might experience those risks. Some assessments use scoring systems to prioritise risks,[19] where others use qualitative descriptions.[20] As in risk identification, the priority of risks will depend on who is involved in the activity – some methods like the Canadian Government's algorithmic impact assessment (AIA) process involve the project team making this assessment,[21] while others like the NHS AIA process involve a participatory panel of patients, doctors and nurses.[22]

3. **Risk mitigation planning:** Assessment processes usually involve compiling and stating planned mitigations for identified risks. If conducted by third parties, as is often the case for human rights impact assessments, there may be recommendations for mitigations.[23] If risks are too great, it is advised that mitigations include the option not to proceed with development or deployment of the system.[24]

19 Treasury Board of Canada Secretariat, 'Algorithmic Impact Assessment Tool' (22 March 2021) https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html accessed 21 February 2023.

20 Ada Lovelace Institute, *Algorithmic impact assessment: a case study in healthcare* (2022) https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/ (Accessed: 19 April 2022);

21 Treasury Board of Canada Secretariat (n 18)

22 Ada Lovelace Institute (n 19)

23 BSR, 'Google Celebrity Recognition API HRIA Executive Summary' (2019) https://www.bsr.org/reports/BSR-Google-CR-API-HRIA-Executive-Summary.pdf accessed 26 February 2023

24 Dillon Reisman and others, 'Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability' (AI Now Institute 2018) https://ainowinstitute.org/wp-content/uploads/2023/04/aiareport2018.pdf?_ga=2.66252014.1929803774.1692709832-1026089197.1692709832&_gl=1*6tp2v7*_ga*MTAyNjA4OTE5Ny4xNjkyNzA5ODMy*_ga_FKQJRSE30T*MTY5MjcwOTgzMS4xLjEuMTY5MjcwOTg2MC4wLjAuMA

4. **Risk monitoring:** Risk monitoring involves planning for the monitoring of particular risks, or to revisit the risk assessment at a particular point in development or deployment. This may be at a lifecycle stage (for example, to revisit before deployment or regular releases), at a regular cadence (for example, annually), or when the system changes significantly. Identifying system change can pose challenges in terms of determining when a sufficient size or scope of change has occurred, so is often combined with a fixed revisitation point.[25]

5. **Communicating risks:** Many risk or impact assessment methods, particularly for the public sector, recommend that findings are published to increase transparency, improve public trust and provide the public or civil society bodies with information on how a system has been tested. These may be in a single repository, as with the Canadian Government's AIA,[26] or alongside details of an AI product, as has been seen in the private sector.[27] However, particularly with private sector and voluntary processes, there is inconsistency in whether findings (or even the information that the risk or impact assessment has been conducted) are made public and, if so, to what level of detail.

Risk and impact assessment methods in practice

Government of Canada algorithmic impact assessment (AIA)[28]

Mandatory risk assessment tool for use by public-sector organisations to determine impact level. Online questionnaire contains 48 risk and 33 mitigation questions on design, algorithm, decision type, impact and data sources.[29]

25 Ada Lovelace Institute (n 19)
26 Treasury Board of Canada Secretariat, 'Open Government Portal' https://search.open.canada.ca/opendata/?collection=aia&page=1&sort=date_modified+desc accessed 21 February 2023
27 Parker Barnes and Andrew Schwartz, 'Celebrity Recognition Now Available to Approved Media & Entertainment Customers' (*Google Cloud Blog*, 30 October 2019) https://cloud.google.com/blog/products/ai-machine-learning/celebrity-recognition-now-available-to-approved-media-entertainment-customers accessed 26 February 2023
28 Treasury Board of Canada Secretariat (n 18)
29 Ibid

Intended for use on external-facing government systems, tools or statistical models used to make an administrative decision about a client (excluding national security).[30]

- **Who is involved?**
  - Assessment conducted by public-sector bodies. Recommended to be completed by a multidisciplinary team with expertise including service recipients, business processes, data and system design decisions.[31]
- **When in lifecycle?**
  - Assessment required twice:
    - 1) Beginning of the design phase of a project.
    - 2) Prior to the deployment of the system.
- **Communication**
  - Completed AIAs are released on the Open Government Portal.[32]
- **Voluntary/mandated?**
  - Mandated for the public sector in Canada.
- **In use/maturity?**
  - Introduced for all systems developed or procured after 1 April 2020.[33]

30 Treasury Board of Canada Secretariat (n 11)
31 Treasury Board of Canada Secretariat (n 18)
32 Treasury Board of Canada Secretariat (n 25)
33 Treasury Board of Canada Secretariat (n 11)

Government of the Netherlands Fundamental Rights and Algorithm Impact Assessment (FRAIA)[34]

A discussion and decision-making tool for government organisations considering developing, procuring, adjusting or using an algorithm. The process looks holistically at possible consequences of use of an algorithm (including inaccuracy, ineffectiveness), with a particular focus on risks of infringing fundamental rights.

- **Who is involved?**
  - Advises that discussion about the various questions should take place in a multidisciplinary team consisting of people with a wide range of specialisations and backgrounds.
- **When in lifecycle?**
  - In decision-making about use of an algorithmic solution (that is, prior to use).
- **Communication**
  - Links to completed FRAIAs included in the Netherlands' public-sector algorithmic transparency standard.[35]
- **Voluntary/mandated?**
  - Currently voluntary. Active discussions in the EU around requiring fundamental rights impact assessments in the forthcoming AI Act are however looking to this model.[36]

34 Ministerie van Algemene Zaken, 'Impact Assessment Fundamental Rights and Algorithms - Report - Government.Nl' (31 March 2022) https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms accessed 21 February 2023.

35 'Dutch Algorithmic Transparency Standard' (Dutch Algorithmic Transparency Standard) https://standaard.algoritmeregister.org/ accessed 17 March 2023.

36 Luca Bertuzzi, 'AI Act: MEPs Want Fundamental Rights Assessments, Obligations for High-Risk Users' (*www.euractiv.com*, 10 January 2023) https://www.euractiv.com/section/artificial-intelligence/news/ai-act-meps-want-fundamental-rights-assessments-obligations-for-high-risk-users/ accessed 21 February 2023.

- **In use/maturity?**

  — In use by some government departments in the Netherlands since 2021.

Technology industry use of human rights impact assessment (HRIA) for AI

Applying human rights impact assessments to AI systems. HRIAs originate in the development sector but are increasingly used to assess the human rights impacts of business practices and technologies.

- **Who is involved?**

  — Typically a third-party brought in to lead the human rights impact assessment, with access to teams at the company, potentially affected stakeholders and independent experts.[37]

- **When in lifecycle?**

  — Varies – many recommend in advanced of system use,[38] but many published examples have been post-deployment.

- **Communication**

  — Sometimes results are published sporadically on company websites.

- **Voluntary/mandated?**

  — Voluntary.

37 Dunstan Allison-Hope, Hannah Darnton and Michaela Lee, 'Google's Human Rights by Design | Blog | Sustainable Business Network and Consultancy | BSR' (30 October 2019) https://www.bsr.org/en/blog/google-human-rights-impact-assessment-celebrity-recognition accessed 27 February 2023.

38 Brandie Nonnecke and Philip Dawson, 'Human Rights Impact Assessments for AI: Analysis and Recommendations' (Access Now 2022) https://www.accessnow.org/cms/assets/uploads/2022/11/Access-Now-Version-Human-Rights-Implications-of-Algorithmic-Impact-Assessments_-Priority-Recommendations-to-Guide-Effective-Development-and-Use.pdf

- **In use/maturity?**

  - Numerous publicised instances.[39]

Microsoft's Responsible AI Impact Assessment[40]

This impact assessment consists of five sections: project overview, intended uses, adverse impact, data requirements and summary of impact. The process and findings are documented in a template. This template includes prompts around fitness for purpose, potential harms and benefits for different stakeholders, as well as questions on fairness, transparency, accountability, reliability and safety. The impact assessment also prompts for goals for mitigation of risks identified.

- Who is involved?

  - Assessment conducted by internal teams in the company, led by one person, with some parts described as requiring teamwork from team members with different expertise.

- When in lifecycle?

  - Early in the system's development, 'typically when defining the product vision and requirements' and before development starts. Additional review and updates annually, or when new intended uses for the system are added, or before expanding system release.[41]

39 Dunstan Allison-Hope, 'Our Human Rights Impact Assessment of Facebook in Myanmar | Blog | Sustainable Business Network and Consultancy | BSR' (5 November 2018) https://www.bsr.org/en/blog/facebook-in-myanmar-human-rights-impact-assessment accessed 17 March 2023.

40 Microsoft, 'Microsoft Responsible AI Impact Assessment Guide' (Microsoft 2022) https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Guide.pdf accessed 27 February 2023.

41 Microsoft, 'Microsoft Responsible AI Standard v2 General Requirements' [2022] Impact Assessment. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf

- **Communication**
  - Not published externally.[42]
- **Voluntary/mandated?**
  - Voluntary.
- **In use/maturity?**
  - In use as standard at Microsoft, resources available for adoption by others.[43]

Council of Europe Human Rights, Democracy and Rule of Law Assurance Framework for AI Systems (HUDERAF)[44]

The HUDERAF is made up of four elements:

1. A 'Preliminary Context-Based Risk Analysis' to establish a high-level sense of risk from the proposed system.
2. A 'Stakeholder Engagement Process' for identifying relevant stakeholders to inform understanding of risk and impact.
3. A 'Human Rights, Democracy and the Rule of Law Impact Assessment' for identifying potential impacts of the system's use.
4. A 'Human Rights, Democracy and Rule of Law Assurance Case' where project teams document their risk and impact assessment processes, the risks identified and mitigation plans.[45]

- **Who is involved?**
  - Conducted by AI project teams, with a process to identify stakeholders and 'facilitate proportionate stakeholder involvement'.

42 Microsoft, 'Responsible AI Principles from Microsoft' (*Microsoft*, June 2022) https://www.microsoft.com/en-us/ai/responsible-ai accessed 21 February 2023.

43 Ibid.

44 David Leslie and others, 'Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems: A Proposal' (2022) http://arxiv.org/abs/2202.02776 accessed 27 February 2023

45 Ibid.

- **When in lifecycle?**

  - Activities across the project lifecycle, beginning with the design phase.

- **Communication**

  - The process recommends that those undertaking it 'publicly communicate HUDERIA findings and impact management plans (action plans) to the greatest extent possible (for example, published, with any reservations based on risk to rights-holders or other participants clearly justified).'[46]

- **Voluntary/mandated?**

  - If the draft Council of Europe Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law is established, for signatory countries this process forms a framework for mandatory compliance with that convention.[47] The impact assessment itself is non-binding.[48]

- **In use/maturity?**

  - Not currently in use.

46 David Leslie and others (n 43)

47 Council of Europe Committee on Artificial Intelligence, Revised zero draft [framework] convention on artificial intelligence, human rights, democracy and the rule of law 2023 https://rm.coe.int/cai-2023-01-revised-zero-draft-framework-convention-public/1680aa193f accessed 27 February 2023

48 European Commission, Recommendation for a COUNCIL DECISION authorising the opening of negotiations on behalf of the European Union for a Council of Europe convention on artificial intelligence, human rights, democracy and the rule of law 2022 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0414 accessed 27 February 2023

Stakeholder impact assessment

A process to document the collaborative evaluation and reflective anticipation of the possible harms and benefits of AI projects. It involves: identifying affected stakeholders; mapping the goals and objectives of an AI project; considering possible impacts on individuals and society; and public consultation. Developed by researchers at the Alan Turing Institute in the UK, particularly with a view to application in the public sector. It is intended to be used alongside other forms of impact assessment applicable in the UK, such as DPIAs and equalities impact assessments (EIAs).[49]

- **Who is involved?**
  - Led by the AI project team, but includes identifying and consulting a wide range of stakeholders, as well as public consultation.
- **When in lifecycle?**
  - At design stage, after development stage (once model has been trained, tested and validated), and iteratively revisited after deployment.
- **Communication**
  - Unclear – includes suggested documentation format that could be published.
- **Voluntary/mandated?**
  - Voluntary.
- **In use/maturity?**
  - In trial, no published examples.

49 David Leslie, '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 2019) https://zenodo.org/record/3240529 accessed 13 January 2020.

UK NHS algorithmic impact assessment (AIA) in healthcare

Application of AIA approach to a data-access process in healthcare. The process involves reflective thinking through possible impacts by project teams, a deliberative process with patients and members of the public and the publication of the final impact assessment. It was developed by the Ada Lovelace Institute in the UK to be trialled with the UK's National Medical Imaging Platform (NMIP), with the data-access process intended to form an accountability mechanism for the consideration and mitigation of risks.[50]

- **Who is involved?**
  - Led by AI project teams, participated in by patients and members of the public, reviewed by an interdisciplinary data-access committee.
- **When in lifecycle?**
  - Before data access, ideally in the early research and design phase, but in practice may be applied to a range of lifecycle points. Intended to be revisited over time.
- **Communication**
  - AIAs of successful applicants for data recommended to be published in one location on the website of the data source.
- **Voluntary/mandated?**
  - Would be required for data access for the NHS NMIP.

50 Ada Lovelace Institute (n 19)

- **In use/maturity?**
  - Planned pilot,[51] has been explored for use in other contexts.[52]

Google Ethical AI team's SMACTR[53]

Framework for internal algorithmic auditing, designed to support end-to-end AI system development. The framework includes five distinct stages: scoping, mapping, artefact collection, testing and reflection.

- **Who is involved?**
  - Designed to be completed by a range of 'key internal stakeholders', such as product teams, management and other stakeholders who have proximity to (or control of) aspects of an AI system (for example, the training data). Suggests that 'diverse expertise' will strengthen the efficacy of the framework.
- **When in lifecycle?**
  - During product development, prior to launch.
- **Communication**
  - Internal transparency and external scrutiny promoted, but no formal requirement for publication of a document containing the audit's results.
- **Voluntary/mandated?**
  - Voluntary.

51 Department of Health and Social Care, 'UK to Pilot World-Leading Approach to Improve Ethical Adoption of AI in Healthcare' (GOV.UK, 8 February 2022) https://www.gov.uk/government/news/uk-to-pilot-world-leading-approach-to-improve-ethical-adoption-of-ai-in-healthcare accessed 26 February 2023

52 Lucas Wright, Winfield Mac and Joshua Clark, 'Implementing Algorithmic Governance: Clarifying Impact Assessments Through Mock Exercises' [2022] SSRN Electronic Journal https://www.ssrn.com/abstract=4349890 accessed 2 March 2023

53 Inioluwa Deborah Raji and others, 'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing' (arXiv, 3 January 2020) http://arxiv.org/abs/2001.00973 accessed 27 March 2023

- **In use/maturity?**

  — Has been put forward for implementation into a medical algorithmic audit process.[54]

US National Institute of Standards and Technology's (NIST) AI Risk Management Framework[55]

- **Who is involved?**

  — 'Different AI stakeholders' – those playing an 'active role' in an AI system's lifecycle, including both developer and vendor organisations, and individuals such as domain experts, designers, compliance experts and advocacy groups.

- **When in lifecycle?**

  — Iterative, continual process designed to be performed throughout different stages of AI lifecycle – but with potential for variance according to individual organisations' schedule / interests.

- **Communication**

  — No formal transparency requirement.

- **Voluntary/mandated?**

  — Voluntary.

- **In use/maturity?**

  — No known published examples.

54 Xiaoxuan Liu and others, 'The Medical Algorithmic Audit' (2022) 4 The Lancet Digital Health e384 https://linkinghub.elsevier.com/retrieve/pii/S2589750022000036 accessed 2 March 2023.

55 Elham Tabassi, 'AI Risk Management Framework: AI RMF (1.0)' (National Institute of Standards and Technology 2023) error: NIST AI 100-1 https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf accessed 3 March 2023.

UK Information Commissioner's Office (ICO) AI and Data Protection Risk Toolkit

The ICO AI and Data Protection Risk Toolkit aims to combine data protection regimes with considerations for AI. It is a spreadsheet of prompts related to UK GDPR and data protection risks at each stage of the AI lifecycle, with accompanying guidance, action recommendations and space for documenting the process.[56]

- **Who is involved?**
  - Targeted at 'organisations using AI'.
- **When in lifecycle?**
  - Adopted/used at different phases within the AI lifecycle.
- **Communication**
  - Unclear – unable to find documentation of use.
- **Voluntary/mandated?**
  - Voluntary, but can help demonstrate compliance with data protection laws.
- **In use/maturity?**
  - Unclear – unable to find documentation of use.

56 ICO, 'AI and Data Protection Risk Toolkit' (19 January 2023) https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/ai-and-data-protection-risk-toolkit/ accessed 27 February 2023.

Example in practice: human rights impact assessment for AI – Google's celebrity recognition API

Google commissioned consultancy BSR to conduct a human rights impact assessment (HRIA) during the product design and development phase of its celebrity recognition application programming interface (API).[57]

The system uses computer vision to enable searching of images for celebrities within a dataset of licensed images of actors, athletes and TV/film celebrities. The celebrity recognition API is made available for selected media or entertainment enterprise customers, to enable them to search and label their image or video libraries.[58]

The HRIA was conducted collaboratively with Google Cloud AI's API product and cross-functional AI principles teams. The methodology outlined was not detailed, but is described as being 'based on the UN Guiding Principles (UNGPs) on Business and Human Rights, including aspects such as consultation with potentially affected stakeholders, dialogue with independent expert resources, and paying particular attention to those at heightened risk of vulnerability or marginalisation'.[59]

The HRIA resulted in the identification of a range of human rights risks relating to privacy, freedom of expression, security, child rights, non-discrimination and access to culture. The assessment recommended actions for Google to take, such as restricting inclusion in the celebrity database to people who are voluntarily the subject of public media attention. The assessment also included recommendations for wider sectoral actors developing similar products, such as participating in efforts to create industry-wide principles of practice on the use of such products. Finally, the assessment included recommendations for users of these kinds of services, such as doing their own HRIA of their particular use of the product.

57 BSR (n 22)

58 Barnes and Schwartz, (2019), *Celebrity Recognition now available to approved media & entertainment customers*, https://cloud.google.com/blog/products/ai-machine-learning/celebrity-recognition-now-available-to-approved-media-entertainment-customers (Accessed: 26 February 2023);

59 BSR, (2019), *Google Celebrity Recognition API HRIA* Executive Summary, https://www.bsr.org/reports/BSR-Google-CR-API-HRIA-Executive-Summary.pdf (Accessed: 26 February 2023);

This HRIA was a single assessment carried out before the system was launched, with no clear requirement for follow-on assessments. It focused on hypothetical uses of the technology, but did not go on to study its actual use post-deployment.

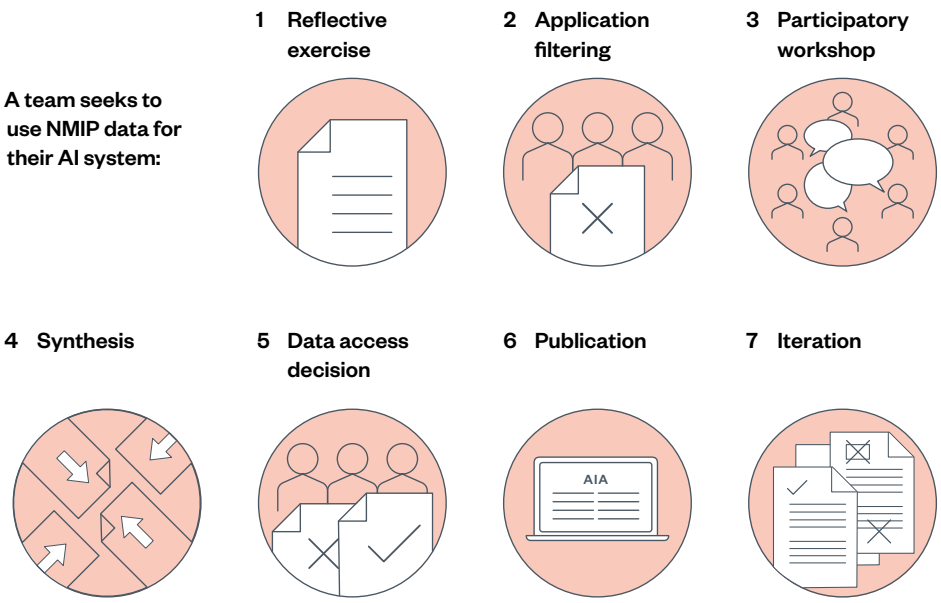
An executive summary report of the HRIA has been made available on the BSR website and is described in a blog post on their website.[60] This is linked to from the blog post announcing the celebrity recognition product.[61] At the time of writing it is not clear how or if the recommendations from the HRIA are communicated to users of the product within the product's interface, or to the celebrities whose images are included in the dataset it matches against.

Example in practice: algorithmic impact assessment in healthcare – NHS medical imaging data access

The UK Government is planning to pilot the use of an algorithmic impact assessment (AIA) as part of the data-access process for National Health Service (NHS) data.[62] The process was developed as part of a research partnership between NHS England's AI Lab and the Ada Lovelace Institute as the first of its kind to explore the potential for AIAs in a real-life healthcare case study: the National Medical Imaging Platform (NMIP).

The AIA process involves a reflexive exercise conducted by research and development teams to identify risks, combined with a participatory workshop with patients, and public involvement to broaden the range of inputs into impact identification. This, along with proposed risk mitigations, is submitted as part of a data-access application to a data-access board, who include the AIA as part of their assessment of whether to grant access. It is recommended that AIAs are made public to communicate risks and the risk assessment process. [63]

60 Allison-Hope, Darnton and Lee, (2019), *Google's Human Rights by Design* | Blog | Sustainable Business Network and Consultancy | BSR, https://www.bsr.org/en/blog/google-human-rights-impact-assessment-celebrity-recognition (Accessed: 27 February 2023);
61 Parker Barnes and Andrew Schwartz (n 26)
62 Department of Health and Social Care (n 50)
63 Ada Lovelace Institute (n 19)



Emerging risk assessment methods for AI in the law

Policymakers worldwide are aiming to incorporate requirements for assessing risks of AI into AI governance regimes and legislation, with risk and impact assessments emerging as common features.

In the EU, the Digital Services Act (2022) requires annual risk assessments for system risks from very large online platforms. With negotiations on the AI Act still in progress, the Parliament's text proposes fundamental rights impact assessments (FRIAs) as a requirement for 'high-risk' AI systems. In the USA, the proposed Algorithm Accountability Act (2022) would mandate businesses that deploy automated decision-making systems and decision processes 'augmented' with AI to produce impact assessments. While this federal bill has yet to be passed, a series of US states like California, New York and Washington have proposed legislation to mandate the use of risk and impact assessments for public sector uses of AI.[64] In Brazil's draft AI legislation, algorithmic impact

64 Sorelle Friedler Engler Suresh Venkatasubramanian, and Alex, 'How California and Other States Are Tackling AI Legislation' (Brookings, 22 March 2023) https://www.brookings.edu/blog/techtank/2023/03/22/how-california-and-other-states-are-tackling-ai-legislation/ accessed 28 March 2023.

assessments (AIAs) must be conducted and made publicly available by providers and for users of 'high-risk' AI systems.[65] In Canada, AIAs are already mandated for public sector agencies.[66]

In the UK, the current language of the draft Online Safety Bill requires online platforms to conduct risk assessments of the prevalence of illegal online content appearing on their services.[67] These assessments will be likely to require platforms to consider risks from AI systems used in content moderation and recommendation systems, which may remove or amplify certain kinds of content to users. Similarly, under the current UK General Data Protection Regulation (GDPR), DPIAs are required for data processing that may be considered 'high risk' under the bill's guidelines.[68]

65 Evangelos Sakiotis, Anna Oberschelp de Meneses and Nicholas Shepherd Quathem Kristof Van, 'Brazil's Senate Committee Publishes AI Report and Draft AI Law' [2023] *Inside Privacy* https://www.insideprivacy.com/emerging-technologies/brazils-senate-committee-publishes-ai-report-and-draft-ai-law/ accessed 28 March 2023.

66 Treasury Board of Canada Secretariat (n 18)

67 'How We Are Approaching Online Safety Risk Assessments' (Ofcom, 14 March 2023) https://www.ofcom.org.uk/news-centre/2023/how-we-are-approaching-online-safety-risk-assessments accessed 28 March 2023.

68 Information Commissioner's Office, 'Data Protection Impact Assessments' (17 October 2022) https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/ accessed 28 March 2023.

Other methods for checking, monitoring and mitigating risks

Identifying and assessing risks alone does not ensure that risks are mitigated or avoided in practice. Many researchers and government agencies have highlighted the need for an ecosystem of AI risk assessment, assurance or audit.[69] This reflects the demand for ways that other actors can check that risks have been appropriately considered and acted on, which in turn relies on methods for monitoring and communicating risks over time. Rather than delegating the task of evaluating risks to a single actor, an ecosystem of risk assessment empowers different actors to conduct and verify risk assessments using a range of different methods.

There are many emerging methods that are being proposed in legislation and experimented with by industry. Some of these methods are already in use (for example, transparency registers), while others are still emerging and are largely unaccounted for in national policy and regulatory proposals (for example, red teaming, documentation standards).

69 Centre for Data Ethics and Innovation, 'The Roadmap to an Effective AI Assurance Ecosystem' (2021) https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance-ecosystem accessed 17 March 2023. Sasha Costanza-Chock, Inioluwa Deborah Raji and Joy Buolamwini, 'Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem', *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2022) https://dl.acm.org/doi/10.1145/3531146.3533213 accessed 17 March 2023. Digital Regulation Cooperation Forum, 'Auditing Algorithms: The Existing Landscape, Role of Regulators and Future Outlook' (2022) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1071554/DRCF_Algorithmic_audit.pdf.

Audit and regulatory inspection

AI auditing is used to refer to a number of different practices that typically involve external scrutiny of an AI system or the processes around it.[70] These practices can be thought of as:

- **Technical audit:** Originating in the computer science community, technical audits adopt the social science practice of an 'audit study' and apply it to algorithmic systems. This form of audit is a narrowly targeted test of a particular hypothesis about a system, usually by looking at its inputs and outputs – for instance, seeing if the system performs differently for different user groups.[71] These methods can be used as standalone audits within companies on their own systems, externally by researchers, journalists or activists, or as part of a compliance audit or regulatory inspection processes.

- **Compliance audit:** This involves checking whether an AI development team has completed processes or met benchmarks sufficient to be compliant with legislation. This form of audit is emerging increasingly in regulation and is anticipated to be conducted by third-party auditors – as a similar process to audit in other fields, such as financial audit.

- **Regulatory inspection:** Inspections are made by regulators, who have powers to investigate and test AI systems for monitoring, suspected noncompliance or verifying claims, such as in legislation on algorithms in social media platforms in the EU's Digital Services Act or the UK's Online Safety Bill.

70 Eticas Consulting, 'Guide to Algorithmic Auditing' (January 2021) https://www.eticasconsulting.com/wp-content/uploads/2021/04/Guide-to-Algorithmic-Auditing-English-Final-ALL-MZ-version-7.pdf accessed 17 March 2023. Ada Lovelace Institute, *Technical methods for regulatory inspection of algorithms in social media platforms* (2021) https://www.adalovelaceinstitute.org/report/technical-methods-regulatory-inspection/ accessed 1 February 2023; Shea Brown, Jovana Davidovic and Ali Hasan, 'The Algorithm Audit: Scoring the Algorithms That Score Us' (2021) 8 Big Data & Society 2053951720983865 https://doi.org/10.1177/2053951720983865; Sasha Costanza-Chock, Inioluwa Deborah Raji and Joy Buolamwini (n 68); Inioluwa Deborah Raji and others, 'Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance' (arXiv, 9 June 2022) http://arxiv.org/abs/2206.04737

71 Ada Lovelace Institute and DataKind UK, 'Examining the Black Box: Tools for assessing algorithmic systems' (2020) https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/ accessed 1 February 2023

- **Sociotechnical assessment:** These processes have been referred to as 'audit', 'sociotechnical audit' or 'internal audit', although in practice they appear more similar to the impact assessment processes described above. They are sometimes carried out in combination with technical auditing approaches or compliance audits.

Each of these interpretations of 'AI audit' can serve an important function. Audits usually come into place after a system is in use, so can serve as accountability mechanisms to verify whether a system behaves as developers intend or claim, whether risk mitigations have been effective and to investigate whether unanticipated impacts have occurred.

Oversight bodies and ethics review committees

Independent oversight bodies have been used to oversee and direct the use of AI, particularly in the public sector, such as the West Midlands Police Data Ethics Committee.[72] In academic AI research, ethics review committees can serve a similar function, and are increasingly being used in industry AI labs.[73] These bodies and committees are typically responsible for: monitoring the actions of project or research teams; reviewing proposals before research or projects are undertaken; and making recommendations, sanctions or decisions about how projects and research teams can develop, use or deploy an AI system.[74] They could be used to review or contribute to risk and impact assessments, and inform or lead decision-making based on identified risks.

Red teaming

Red teaming is an approach originating in computer security. It describes approaches where individuals or groups (the 'red team') are tasked with looking for errors, issues or faults with a system, by taking on the role of

72 West Midlands Police & Crime Commissioner, 'Ethics Committee' (West Midlands Police & Crime Commissioner) https://www.westmidlands-pcc.gov.uk/ethics-committee/ accessed 27 February 2023.

73 Ada Lovelace Institute, *Looking before we leap: Expanding ethical review processes for AI and data science research* (2022) https://www.adalovelaceinstitute.org/report/looking-before-we-leap/ accessed 27 February 2023

74 Ada Lovelace Institute, AI Now Institute and Open Government Partnership, 'Algorithmic accountability for the public sector' (2021) https://www.adalovelaceinstitute.org/report/algorithmic-accountability-public-sector/

a bad actor and 'attacking' it. In the case of AI, it has increasingly been adopted as an approach to look for risks of harmful outputs from AI systems.[75]

For instance, AI research lab Anthropic recruited online workers to probe an AI chatbot, to try to discover and measure harmful outputs from language models, such as the chatbot recommending violent acts, or expressing hateful and/or racist statements.[76] However, this approach currently lacks standards and norms. There are risks to the workers recruited to red teams, particularly in red-teaming scenarios at scale, where there is a skew towards lower-paid crowd-workers.[77]

Safety checklists

Safety checklists have a history in engineering and manufacturing, but have also been seen applied to safety in medical settings and aviation.[78] Checklists can be used both to prompt or check completion of actions, but also to prompt discussion of risks.[79] For AI, safety checklists have been proposed to help teams consider a range of risks and 'check' that best practices have been followed across the AI lifecycle.[80]

The European Commission's High Level Expert Group on AI has created an 'Assessment List for Trustworthy Artificial Intelligence' that uses both a written and interactive checklist of prompts on a range of rights-based issues and expected actions.[81] Safety

75 Miles Brundage and others, 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims' (arXiv, 20 April 2020) http://arxiv.org/abs/2004.07213 accessed 13 February 2023.

76 Deep Ganguli and others, 'Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned' (arXiv, 22 November 2022) http://arxiv.org/abs/2209.07858>accessed 10 February 2023.

77 Mark Diaz and others, 'CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation', *2022 ACM Conference on Fairness, Accountability, and Transparency* (2022) http://arxiv.org/abs/2206.08931 accessed 17 March 2023.

78 Michael A Madaio and others, 'Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI', *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery 2020) https://dl.acm.org/doi/10.1145/3313831.3376445 accessed 17 March 2023.

79 Michael A Madaio and others, 'Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI', *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery 2020) <https://dl.acm.org/doi/10.1145/3313831.3376445> accessed 17 March 2023.

80 Ibid.

81 High-Level Expert Group on Artificial Intelligence, 'Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment' (2020) https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment accessed 26 February 2023.

checklists could be used to set expectations and monitor completion of a range of risk assessment and mitigation tasks for AI systems, or as a starting point for a list of risks to consider. However, to date they offer little detail on how to assess, weigh or mitigate those risks, and so would need to be combined with other activities.

Model and dataset documentation methods

Good documentation of AI systems can help support appropriate use. Approaches to standardising documentation of how a system works include model cards (which document information about an AI system's architecture, testing methods, and intended uses)[82] and datasheets (which document information about a dataset, including what kind of data is included, how it was collected, and how it was processed).[83] These documentation methods often include prompts asking developers of an AI system or dataset to consider and document the potential risks it may pose.

These tools recognise that AI models and datasets are often used by downstream deployers in an AI supply chain, who will need to understand technical details, development and data collection contexts, and risks that may only be known to upstream developers of that system or dataset. Model cards can include details of findings from risk assessments, while both model cards and datasheets could be useful in informing risk and impact assessments for AI systems that implement or use documented models or datasets.

Transparency registers

Where model cards and datasheets often focus on actors in the AI supply chain, there are also frequent calls for transparency about AI systems and their risks to end users, impacted groups and the wider public. In particular in the public sector, registers of AI systems have been

82 Margaret Mitchell and others, 'Model Cards for Model Reporting', *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019) http://arxiv.org/abs/1810.03993 accessed 1 February 2023.

83 Ben Hutchinson and others, 'Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2021) https://dl.acm.org/doi/10.1145/3442188.3445918 accessed 17 March 2023; Timnit Gebru and others, 'Datasheets for Datasets' (December 2021) https://m-cacm.acm.org/magazines/2021/12/256932-datasheets-for-datasets/abstract accessed 27 February 2023.

proposed as a way to collate documentation about systems in use and make it available to the public.

In the UK, the Algorithmic Transparency Standard and the Algorithmic Transparency Recording Standard Hub are a step towards this, trialling a standard method for reporting on public-sector AI systems.[84] The Transparency Standard includes a section on risks, and requests for outputs from impact assessments such as algorithmic impact assessments (AIAs), or data privacy impact assessments.

Registers have been trialled at city level in Amsterdam, Antibes, Barcelona, Brussels, Eindhoven, Lyon, Mannheim, Nantes, Ontario, Rotterdam, Sofia and Helsinki.[85] In Chile, pilots are moving towards a General Instruction on Algorithmic Transparency by the Chilean Transparency Council that will mandate what information about public-sector systems is to be made available.[86]

84 Ada Lovelace Institute, AI Now Institute, and Open Government Partnership (n 73); Natalia Domagala and Hannah Spiro, 'Engaging with the Public about Algorithmic Transparency in the Public Sector' (*Centre for Data Ethics and Innovation Blog*, 21 June 2021) https://cdei.blog.gov.uk/2021/06/21/engaging-with-the-public-about-algorithmic-transparency-in-the-public-sector/ accessed 1 February 2023.

85 Ada Lovelace Institute, AI Now Institute, and Open Government Partnership (n 73); 'Algorithm Register - Algorithmic Transparency Standard' <https://www.algorithmregister.org/> accessed 27 February 2023.

86 Consejo para la Transparencia, 'Consejo para la Transparencia y Universidad Adolfo Ibáñez lideran piloto en organismos públicos para inédita normativa en transparencia algorítmica de América Latina' (*Consejo para la Transparencia*, 17 October 2022) https://www.consejotransparencia.cl/consejo-para-la-transparencia-y-universidad-adolfo-ibanez-lideran-piloto-en-organismos-publicos-para-inedita-normativa-en-transparencia-algoritmica-de-america-latina/ accessed 27 February 2023; 'Algoritmos Públicos - GobLab UAI' https://www.algoritmospublicos.cl/ accessed 27 February 2023.

Enabling an ecosystem of risk assessment

The relatively immature and fragmented landscape of AI risk-assessment methods presents the UK Government with an opportunity to lead in the development and standardisation of these practices. Some technology companies like Google and Microsoft have experimented with some of these methods in anticipation of forthcoming regulation requiring their use and it is likely that many UK technology companies are also considering the adoption of these mechanisms. There is an urgent need for Government action to create a standardised method of assessment in coordination with other national bodies developing these methods.

What could the Government do to create an effective assessment ecosystem?

- **Create incentives for companies and third parties to assess risks from AI systems.** Methods for AI risk assessments are not yet mainstream or default in AI system development or deployment. In the private sector, adoption is challenging to measure due to lack of public reporting. In the public sector there is some adoption or trialling of reporting mechanisms (such as the UK's Algorithmic Transparency Standard, which includes requests for information about potential risks, and for links to results of impact assessments), but it is still sporadic. Many jurisdictions are looking to regulation to increase incentives to assess societal risks – and many actors are calling for mandates– from Canada's mandated algorithmic impact assessment (AIA) for public sector agencies,[87] to fundamental rights impact assessments (FRIAs) in the Netherlands. In other locations, such as Chile, there are moves to mandate transparency reporting which, if they include information about risk or risk assessment, could

87 Emanuel Moss and others, 'Governing with Algorithmic Impact Assessments: Six Observations' (24 April 2020) https://papers.ssrn.com/abstract=3584818 accessed 27 February 2023.

help incentivise risk assessment processes.[88] Other incentives could include introducing requirements as part of data-access processes and procurement requirements in the public sector, as well as strengthening government or regulatory advice around best practice in AI or what to look for when procuring AI systems. This might also include establishing prizes or competitions around risk assessment methods or trials as direct drivers of examples of good practice.

- **Case studies of risk assessment methods in practice.** There are still few published examples of risk assessments of real AI systems, which can make it hard to compare or evaluate risk assessment methods, or to understand good practice. To improve this, more published case studies of algorithmic risk assessments are needed, documenting how the process changed or shaped the design, development and outcomes. This could be aided by collaboration with independent researchers and civil society to help conduct or evaluate this work.[89]

- **Standards for assessing risks.** There is presently no consensus on standards for risk and impact assessments. This is understandable as these methods are still being trialled and developed, but until standards are agreed, there remains a risk that any AI developer can claim to have conducted an assessment with no guarantee or indication for the public as to what it entailed. There has also been discussion about mandated standards: if FRIAs are brought in as part of the EU AI Act, there has been debate about how the details of these assessments would be established. Some have suggested technical standards bodies could develop these processes, but there is concern that these lack the required mix of skills and accountability to affected communities.[90]

- **Domain or sector-specific guidance on societal risks.** The development of guidance in sector-specific areas – such as healthcare, social care, workforce management or recruitment – could help complement broader guidance or standards for using these methods. While many of the approaches examined above include prompts for different forms of societal risks, they will inherently be

88 Consejo para la Transparencia (n 85)
89 Ada Lovelace Institute and DataKind UK (n 70)
90 Ada Lovelace Institute, *Inclusive AI governance: Civil society participation in standards development* (2023) https://www.adalovelaceinstitute.org/report/inclusive-ai-governance/

limited by the expertise of those designing the specific assessment processes. Broadly applicable AI risk and impact assessment methods could be tailored to sectors through additional guidance, prompts or adaptations of methods, as has already been seen in the case of healthcare, with the adaptation of the SMACTR framework, or AIAs to healthcare-specific use cases.

- **Skills and roles in the technology sector.** The technology sector will need teams, roles and staff with the skills to conduct risk and impact assessments. In particular, many methods involve identifying and coordinating diverse stakeholders, and the use of participatory or deliberative methods which are not currently widespread in the technology sector, but are more established in other domains such as policy, design, academic sociology and anthropology.[91] Some of the skills of user research, which is more widely applicable in technology development, may be transferable to these methods, though the focus and intention of the role would be different.

- **Regulatory capacity.** This will be important to support an ecosystem of risk assessment, and to deliver monitoring and investigation functions that help ensure the mitigation of risks over time.[92] In the UK, for instance, some regulators have had established capacity for this over a long period, such as the Competition and Markets Authority (CMA) and others such as Ofcom, which have recently been expanding to take on new regulatory responsibilities over online safety. However, some regulators that have expertise that is well-suited to considering societal risks are currently under-resourced to tackle questions of AI. They would need both increased resources and to be empowered to investigate risks and harms from AI systems.

- **Empowering third-party risk and impact assessors.** Many of the most well-known and significant AI risk assessments to date have been conducted by third-party civil society groups, academics and companies that evaluate a system's impacts without the permission of the company.[93] For example, evaluations of bias and the transferring

91 Sasha Costanza-Chock, Inioluwa Deborah Raji and Joy Buolamwini (n 68)

92 Ada Lovelace Institute, Regulate to Innovate (2021) https://www.adalovelaceinstitute.org/report/regulate-innovate/ accessed 1 February 2023; Sasha Costanza-Chock, Inioluwa Deborah Raji and Joy Buolamwini, (n 68); Digital Regulation Cooperation Forum (n 68); Mhairi Aitken and others, 'Common Regulatory Capacity for AI' (Zenodo 2022) https://zenodo.org/record/6838946

93 Sasha Costanza-Chock, Inioluwa Deborah Raji and Joy Buolamwini (n 68)

of sensitive medical data by Facebook (Meta) have largely been conducted by third-party researchers at organisations like The Markup.[94] Third-party assessors are independent, and can bring local context or consideration to the evaluation of a system's impacts. However, third parties often lack access or information about emerging AI systems, and may not be well resourced to conduct these kinds of assessments. In emerging regulation, governments can empower third parties to have greater mandates to access critical data or technical information about an AI system that can enable this kind of assessment of risks.

94 Todd Feathers and others, 'Facebook Is Receiving Sensitive Medical Information from Hospital Websites – The Markup' (16 June 2022) https://themarkup.org/pixel-hunt/2022/06/16/facebook-is-receiving-sensitive-medical-information-from-hospital-websites accessed 27 March 2023.

Further questions

This section briefly outlines further questions or opportunities for research.

- What kinds of risk assessment methods would work well for the UK's public sector?
- What specific kinds of support do third-party assessors need to better conduct their assessments?
- How often should companies developing or using AI undertake risk assessments?
- Drawing on lessons from other fields, how effective is making risk and impact assessments publicly available at improving transparency and public trust?
- What kinds of standards and professional practices are needed to create an ecosystem of risk assessment of AI?
- Should the public and private sectors have the same obligations to undertake risk assessments of AI systems?
- How can risk assessment methods involve those affected or impacted by AI systems?
- What lessons can the UK learn from other national risk assessment requirements, such as the Netherlands' fundamental rights algorithmic impact assessment (FRAIA) and Canada's algorithmic impact assessment (AIA)?
- What role can risk assessments play in the public-sector procurement process in the UK? What methods would work best?

Methodology

This report surveys approaches for assessing risks that AI systems pose for people and society – in practice, on the ground within AI project teams and in emerging legislation.

The findings of this report are based on a desk-based review and synthesis of grey and academic literature on approaches to assessing AI risk. Relevant literature was identified through keyword searching of academic and grey literature databases, and snowball sampling through the community of practitioners working on AI risk management.

The report is also informed by policy analysis of draft legislation related to governance of AI and algorithmic systems, primarily in a UK and European context, and focused on requirements for anticipating risks or impacts of such systems. This analysis is limited to legislation drafted – or with documentation available – in English, and the research team acknowledges that it would benefit from further work considering wider linguistic, geographic and political contexts.

Partner information and acknowledgements

This report was authored by Jenny Brennan, with substantive contributions from Lara Groves, Elliot Jones and Andrew Strait.

This work was undertaken with support via UKRI by the DCMS Science and Analysis R&D Programme. It was developed and produced according to UKRI's initial hypotheses and output requests. Any primary research, subsequent findings or recommendations do not represent DCMS views or policy and are produced according to academic ethics, quality assurance and independence.

About the Ada Lovelace Institute

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.

The mission of the Ada Lovelace Institute is to ensure that data and AI work for people and society. We believe that a world where data and AI work for people and society is a world in which the opportunities, benefits and privileges generated by data and AI are justly and equitably distributed and experienced.

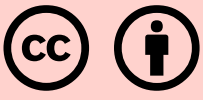
We recognise the power asymmetries that exist in ethical and legal debates around the development of data-driven technologies, and will represent people in those conversations. We focus not on the types of technologies we want to build, but on the types of societies we want to build.

Through research, policy and practice, we aim to ensure that the transformative power of data and AI is used and harnessed in ways that maximise social wellbeing and put technology at the service of humanity.

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.

Find out more:

Website: Adalovelaceinstitute.org
Twitter: @AdaLovelaceInst
Email: hello@adalovelaceinstitute.org



Preferred citation: Ada Lovelace Institute. *AI assurance? Assessing and mitigating risks across the AI lifecycle* (2023).
https://www.adalovelaceinstitute.org/report/risks-ai-systems/

ISBN: 978-1-7392615-5-9