Shaping the ethical dimensions of smart information systems– a European perspective (SHERPA)

Deliverable No. 5.8

Artificial Intelligence Impact Assessments

A Systematic Review

31.10.2021

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Document Control

Deliverable	D5.8
WP/Task Related	WP 5 / T5.8
Delivery Date	31.10.2021
Dissemination Level	PUB
Lead Partner	DMU
Authors	Bernd Carsten Stahl (DMU) Josephina Antoniou (UCLanCY) Nitika Bhalla (DMU) Laurence Brooks (DMU) Philip Jansen (UT) Blerta Lindqvist (FSC) Alexey Kirichenko (FSC) Samuel Marchal (FSC) Rowena Rodrigues (TRI) Nicole Santiago (TRI) Zuzanna Warso David Wright (TRI)
Reviewers	peer review by authors
Abstract	Artificial intelligence (AI) is expected to produced impacts that are highly beneficial, but it also raises concerns about undesirable ethical and social consequences. There is an array of activities that aim to address these undesirable consequences, ranging from proposals for regulation such as the EU AI Act to ethics guidelines to design methodologies, professional guidance or standardisation. One option that is increasingly explored is to develop impact assessments specifically geared for the needs of AI. A number of such AI impact assessments (AI-IAs) have already been proposed. This document undertakes a systematic review of these AI-IAs with the aim of identifying whether there are common themes and approaches. This research is important to establish a baseline for AI-IAs that can help organisations identify AI-IAs that are most relevant to their needs and that can serve as a measure for legislators and regulators to determine the role that AI-IAs can play in the governance of the broader AI ecosystem.
Key Words	AI, impact assessment, systematic review, AI governance

Revision History

Version	Date	Author(s)	Reviewer(s)	Notes

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

1	31.10.2021	see above	co-author peer review	

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Table of Contents

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Executive Summary

This deliverable describes the review of AI impact assessments (AI-IAs) undertaken as part of the SHERPA project impact acceleration activities.

The identification of AI-IAs was undertaken by following several established methods. This started with a search of four relevant databases (Scopus, ACM, ISI, IEEE) to identify documents that had undergone academic peer review. In addition, web searches were undertaken using three search engines (Google, Duckduck Go, Bing). Finally, the consortium undertook snowball and peer searches, which included a mail-out to experts and relevant email lists to identify documents. Following the application of inclusion and exclusion criteria a population of 37 AI-IAs remained which was fully analysed.

The analysis covered the following topics in each of the documents:

- Purpose
- Scope
- Organisational context
- Issues (to be identified in the AI-IAs)
- Timeframe
- Process and methods
- Transparency
- Challenges

The analysis shows that there is a certain level of coherence between the AI-IAs, but there are also significant differences. Numerous challenges remain, e.g. conceptual consistency, status of the AI-IA, the possibility of misuse and ethics washing and their role in the broader societal discussion of AI and ethics.

We therefore develop a baseline for AI-IAs that can help decisionmakers decide which AI-IA they wish to use or to develop bespoke ones where needed.

The conclusion, in line with the other work of the SHERPA project is that AI-IAs have a potentially important role to play. They need to be understood, however, in the context of rapidly evolving AI innovation ecosystems. In order to be useful in this context, they need to be continually reviewed and revised and integrated with other ways of addressing the ethics of AI, such as standardisation, education, regulation and they need to be embedded in social and organisational processes.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Background

This deliverable (D5.8 : Artificial Intelligence Impact Assessment - A systematic review) was added to the SHERPA DoA as part of the amendment procedure undertaken in spring 2021.

The rationale for this addition was that the consortium included the following recommendation in its set of recommendations: "Develop baseline model for AI impact assessments" (https://www.project-sherpa.eu/ai-impact-assessment/). This recommendation was arrived at on the basis of research and stakeholder consultation. It forms part of the group of recommendations that fall under the "knowledge base" headings, i.e. those recommendations that are meant to ensure that an AI ecosystem has the knowledge and capacity to maintain and develop this knowledge that is required for the ecosystem to act in ways that are conducive to human flourishing.

The SHERPA project had done significant work on most of its other recommendations and could provide substantive input into them (e.g. regulatory proposals, ethics by design, standardisation). However, the AI-IA component had not been part of the original proposal and thus was not subject to specific work by the consortium. The addition of the new task 5.8 and this resulting deliverable were meant to address this limitation.

The description of he task in the DoA is as follows:

> "This task aims to establish how an impact assessment for AI should be designed, so that it can address ethical and human rights concerns. In order to do this, the task undertakes a systematic review of impact assessments that are of relevance to AI to understand what counts as good practice in impact assessment and how such good practice can be applied to AI. Following the academic literature review, the task will conduct a series of snowball and peer searches to provide a comprehensive account of available AI impact assessment models, tools and templates. These documents will be systematically analysed to identify good practice and minimum requirements for impact assessments to be suitable to address ethical and human rights issues of AI." (Task description T 5.8)

In order to ensure that the work undertaken here is publicly available and visible, the consortium decided to approach this task from the outset as an academic publication project. The remainder of this deliverable therefore takes the form of an academic paper that describes our work, findings and recommendations. Style, substance and formatting are geared towards this aim of having it published in a leading journal. The actual submission of the paper will follow the submission of this deliverable and most likely further development of the paper. It is therefore likely that the published version will differ from the text included here, due to changes introduced during the peer review process.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Artificial Intelligence Impact Assessments

Abstract

Artificial intelligence (AI) is expected to produced impacts that are highly beneficial, but it also raises concerns about undesirable ethical and social consequences. There is an array of activities that aim to address these undesirable consequences, ranging from proposals for regulation such as the EU AI Act to ethics guidelines to design methodologies, professional guidance or standardisation. One option that is increasingly explored is to develop impact assessments specifically geared for the needs of AI. A number of such AI impact assessments (AI-IAs) have already been proposed. This document undertakes a systematic review of these AI-IAs with the aim of identifying whether there are common themes and approaches. This research is important to establish a baseline for AI-IAs that can help organisations identify AI-IAs that are most relevant to their needs and that can serve as a measure for legislators and regulators to determine the role that AI-IAs can play in the governance of the broader AI ecosystem.

Keywords: AI, impact assessment, systematic review, AI governance

Introduction

Artificial intelligence (AI) is expected to revolutionise many aspects of our lives, drive efficiency in organisations, improve processes and make better use of resources. Its significant potential economic and social benefits are, however, counterbalanced by potential disadvantages. There are concerns about consequences for individuals, for example when biased systems promote unfair discrimination[1], affect their access to social services[2], but also about consequences for groups and society, for example, differential profiling and treatment of groups[3], political interference[4] or when AI leads to concentration of wealth and power[5], thus exacerbating existing inequalities.

The discussion of how benefits and disadvantages of AI can be understood and balanced covers a range of stakeholders and disciplines. Proposals for proactively addressing possible problems range from ethical guidelines[6] and codes and professionalism[7] to organisational risk management[8], regulatory actions[9], the strengthening of human rights [10,11] and the creation of new institutions[12,13]. These different possible responses to possible negative ethical and human rights consequences of AI need to be seen in conjunction. It is unlikely that any one of them individually will be able so overcome these issues, but collectively they promise ways of understanding and engaging with these issues. There are frequent references to 'AI ecosystems', in particular in the policy-oriented literature [14–16] which indicate a realisation that a holistic approach will be required.

However, even when using a holistic approach, the question of a suitable starting point remains. When a new AI system transitions from the conceptual stage to design, development and deployment, its technical features, organisational and societal uses become increasingly clear which then calls for critical reflection of the balance between benefits and downsides. One possible avenue to understand possible problems early in the system life cycle and put in place appropriate mitigation measures is to undertake impact assessments for AI. Impact assessments are not a new idea and have a long history in the form of social impact assessment[17], environmental impact assessment[18], human rights impact assessments [19] as well as more topic specific impact assessments such as privacy impact assessments [20,21], data protection impact assessments [22] or ethics impact assessment[23].

The idea to apply an impact assessment approach has been proposed in the academic literature[24,25] and has found resonance in national policy[26] international bodies, such as the European Data Protection Supervisor[27], the European Fundamental Rights Agency[28] and UNESCO[14]. Such an impact assessment could

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

be supported and/or mandated by a relevant regulatory framework, such as the one proposed by the EU[29]. It could help organisations understand their obligations by providing a basis for their risk assessment of AI[8] and regulators to ensure that organisations address issues appropriately. It could be a crucial component in the AI ecosystem that ensures that ethical and human rights aspects are taken into consideration and dealt with appropriately and thereby contribute to well-deserved trust in these technologies.

While there are some initial proposals for an AI impact assessment (AI-IA), there is at present no rigorous academic research on AI-IAs. We therefore ask in this paper what the current landscape of AI-IAs looks like with a view to understanding whether dominant themes and topics can be identified. This will allow for the description of a baseline AI-IA that can inform the development of specific AI-IAs but also of organisational, national and international AI policy.

Methodology

We undertook a systematic review of AI-IAs. Systematic literature reviews constitute a well-described and well-understood research method[30]. Rowe[31], following Schwarz et al.[32] suggests that literature reviews can have several goals: to summarize prior research, to critically examine contributions of past research, to explain the results of prior research found within research streams and to clarify alternative views of past research. In our case we aim to establish a baseline of existing impact assessments with a view to establishing good practice for future AI-IAs.

While methodologies for systematic literature reviews are well-established, there are different ways that a systematic literature review can be undertaken in terms. The main type of input data we were interested in was text describing existing impact assessment with likely relevance to AI. The challenge we face is that impact assessment are practice-oriented documents that can originate from professional bodies, companies, standardisation bodies, regulatory bodies. There are no comprehensive databases that collect such work. We therefore undertook a multi-pronged approach to identify relevant impact assessments by looking at three bodies of work: a) a systematic review of the academic literature, b) general internet search and c) snowball and peer searches. The data collection protocol follows precedent on systematic reviews of ethical issues in IT[33] rather than meta-review methods in the biomedical science[34] which is based on methodological assumptions (quantitative data, representativeness of samples etc.) that do not hold for the qualitative data of the AI-IAs we were interested in.

Key questions of relevance to all three streams of identifying AI-IAs relate to the two core concepts of AI and impact assessment. Our focus is on general applicability and visibility, which is why we used the concept of AI, using search terms "artificial intelligence" and "AI". We added the term "algorithm*" as several early examples of AI-IAs used this term, as in "algorithmic impact assessment"[35–37]. We only included documents that proposed impact assessments of AI. We encountered many examples of impact assessment that made use of AI, e.g. for environmental impact assessments[38,39] but excluded these from the analysis. We furthermore excluded documents that may serve as part of AI-IA but that have a broader scope, such a the recent IEEE Standard 7000-2021[40] that focuses on systems development more broadly the CEN / CENELEC CWA 17145[23] that explores ethics assessment for research and innovation more broadly.

The second general conceptual choice we made was to focus on AI-IAs and exclude documents that only discuss AI-IAs. The International Association for Impact Assessment suggest that an impact assessment is "a structured a process for considering the implications, for people and their environment, of proposed actions while there is still an opportunity to modify (or even, if appropriate, abandon) the proposals"[41]. Such impact assessments are meant to be applied to decision making. We therefore only included documents that provided clear evidence of being intended as AI-IAs, e.g., by detailing required processes, scoring criteria or decision relevance. In practice the dividing line between AI-IAs and texts about them was

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

not always clear, leading to case-by-case discussions and decisions on inclusion / exclusion by the consortium.

The search of the academic literature used four databases: IEEE, Scopus, ISI and ACM, covering both general academic literature and key databases in the AI / computer science field. Realising that most current AI-IAs are practice-oriented and not published in academic outlets, we undertook searches using three search engines (Google, Bing, Duckduckgo). In each case we checked the top 50 hits individually to see whether they contained AI-IAs. Finally, we undertook a set of snowball searches and sought peer input. Snowball searches were triggered by references in any of the other search methods. Realising that there may be AI-IAs in use or development in organisations that are not (yet) publicly shared, we directly contacted 242 organisations whom we knew to be active in the AI field. We also sent out a request for contributions to eight email lists. All of these contacts were pointed to a web-based survey page where we shared the AI-IAs we had already identified and asked for further suggestions.

The following figure represents the logic of our method of identifying AI-IAs:

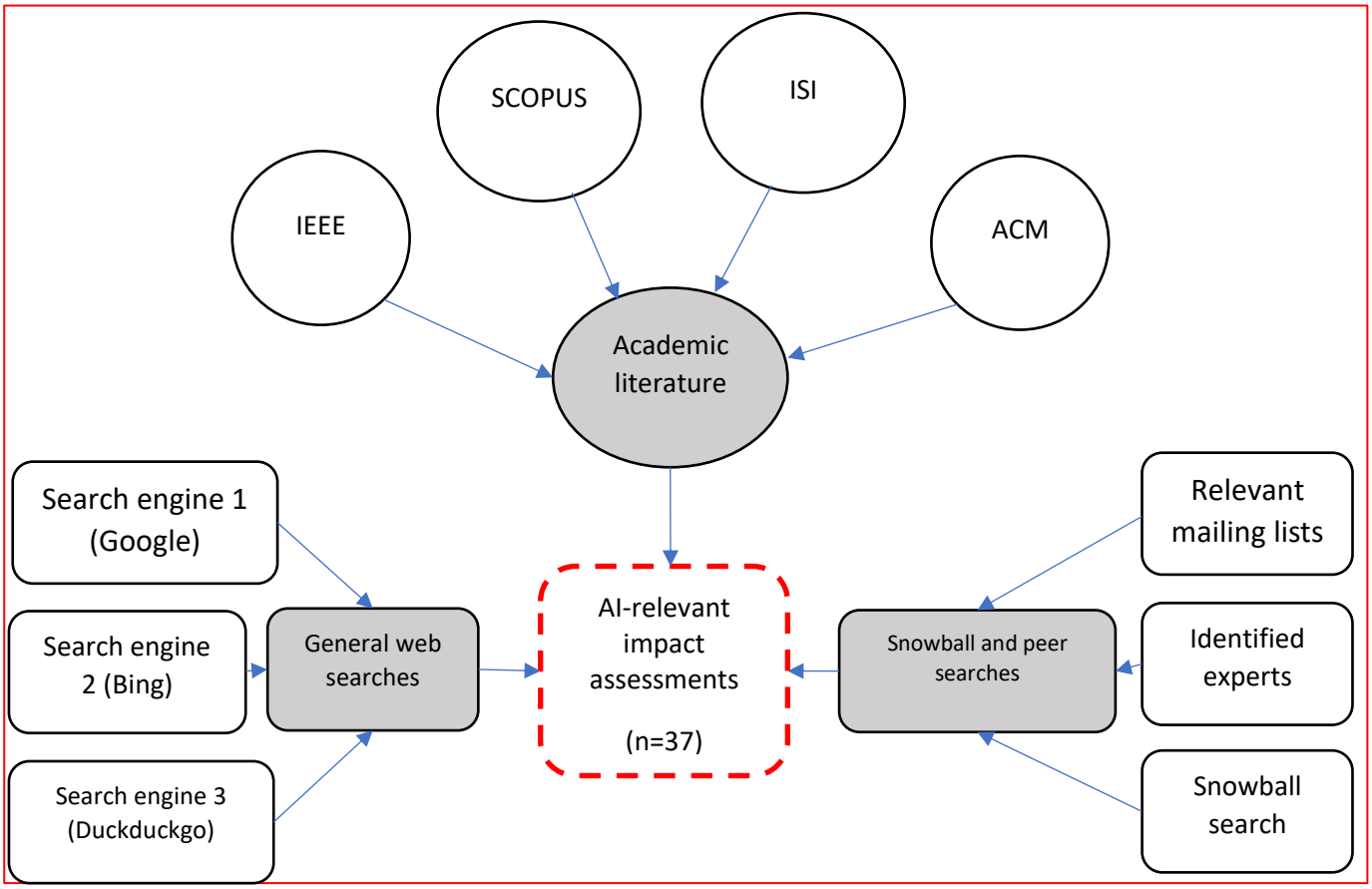

Figure 1: graphical representation of the methodology employed to systematically identify AI-IAs

The method of identifying documents as described in figure 1 led to the identification of 181 unique documents to be potentially included in the analysis, after duplicates were removed. This initial sample then underwent a check using the exclusion criteria described above. The application of the exclusion criteria led to the exclusion of approximately ¾ of the sample. In most cases they were excluded because they used AI in other types of impact assessment, e.g. environmental impact assessment, or because they discussed AI-IAs but did not provide practical guidance on how to undertake them. The remaining 43 documents were included in the analysis as described below. During the analysis another 6 documents were excluded, as more detailed reading revealed that they fell under the exclusion criteria. The final set of

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

AI-IAs that were fully analysed are publicly available via a Zotero group library (https://www.zotero.org/groups/4042832/ai_impact_assessments).

Process Stage	**Numbers of resulting texts**

Identification of texts	Sources:
	• academic literature
	  ○ IEEE = 16
	  ○ Scopus = 81
	  ○ ISI = 17
	  ○ ACM = 8
	• web searches = 47
	• snowball and peer searches = 12

Initial sample (all sources minus duplications)	181
Excluded based on exclusion criteria	43
Excluded during coding exercise	6
Included in final set	37

Table 1: Overview of sample, inclusion and exclusion

The analysis of the AI-IAs was undertaken collectively using the qualitative data analysis software tool NVivo Server version 11. In order to ensure consistency of analysis, an analysis framework was constructed using thematic analysis principles [42,43]. We started with a set of top-level analysis nodes that were defined according to a general view of likely content of an impact assessment. We hypothesised that an impact assessment could usefully include the following components: The analysis of the AI-IAs was undertaken collectively using the qualitative data analysis software tool NVivo Server version 11. In order to ensure consistency of analysis, an analysis framework was constructed using thematic analysis principles [42,43]. We started with a set of top-level analysis nodes that were defined according to our expectations of likely content of an impact assessment. We hypothesised that an impact assessment could plausibly include the following components:

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

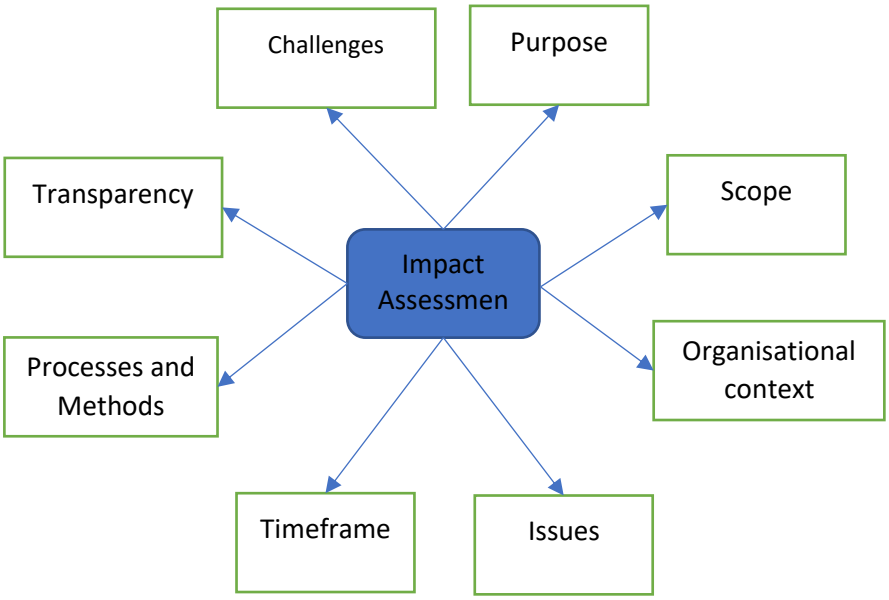

Figure 2: Main analysis topics

This figure embodies our assumptions about AI-IAs as follows: We assumed that they would state a purpose for an IA. They could specify their scope and the organisational context in which they are undertaken. We expected to find a description of the issues they are likely to face and the timeframe in which the AI-IA is to be undertaken. We assumed that there would be a specification of processes and methods used as well as sanction for failure to do the AI-IA. We expected there to be a reference to how transparent the AI-IA itself would need to be and a general description of challenges that can arise during the AI-IA.

A pilot data analysis was undertaken on two high-profile documents that that constitute AI-IAs[35,44]. This allowed us to check the original nodes and to ensure inter-coder reliability. The kappa-coefficient was determined to be between 0.648 and 0.792 in a pairwise comparison between the lead coder and team members. A Kappa of between 0.40 and 0.75 is seen as fair to good agreement with a Kappa over 0.75 counting as excellent[45]. Being satisfied that inter-coder reliability was sufficient, the project team met on a two-weekly basis to discuss findings and agree on the development of the coding scheme on the basis of insights generated during data analysis.

Findings

The final set of 37 documents constitutes a heterogenous mix. Some of the AI-IAs are traditional documents published by individuals. Several of them do not show individual authors but are attributed to organisations or public bodies. Some implement the assessment activities in their presentation or structure, for example when they are implemented as interactive online tools[36] or where they point to supplementary material to be used for assessment purposes[46].

The findings of our analysis are structured along the main analysis nodes as indicated in *Figure 2* above and reflected in the structure of this section. The following figure shows a word cloud generated from the text of all AI-IAs, giving an indication of key terms:

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Figure 3: Word Cloud of all AI-IAs

The analysis followed the nodes as indicated in *Figure 2* above. In total we coded 3975 references to 44 nodes with the distribution of codes distributed as shown in the following figure:

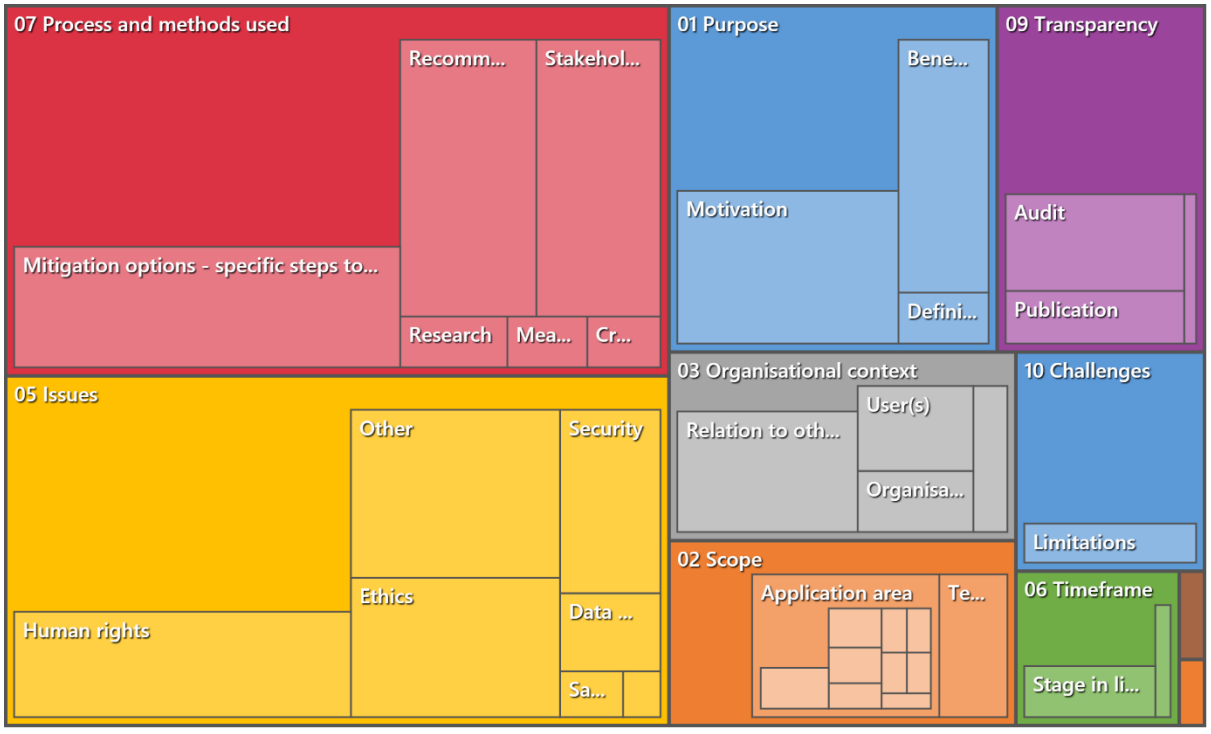

Figure 4: Distribution of most widely used codes during the analysis

The remainder of this section presents our findings.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Purpose

Most of the AI-IAs we analysed state their motivation and purpose, which often included a definition of the AI-IA they offer. The motivation for creating an AI-IA can start with current gaps. These include that purely technical assessments are insufficient[19], a lack of hard law and established quality assessment methods[47]. The motivation for the creation of the assessment then covers a number of intended outcomes, such as safeguarding the benefits of AI[44], understanding their impacts [48–50], assessing systems acceptability[35], and overall promoting trustworthy AI[24,44]. These goals are intended to be achieved or promoted by a number of processes that motivate the development of AI-IAs, such as improvements of communication[51], provision of specific methodologies[52] which promote good practice, e.g., in data protection, [53] and more broadly supporting reflection[51].

The AI-IA documents we surveyed suggest that undertaking such an assessment can have numerous benefits which can be split in functional, organisational, individual and social benefits. Functional benefits are those that suggest that undertaking an assessment will lead to better AI systems. AI-IAs aim to achieve this by pointing to known weaknesses, such as biases in machine learning, strengthening accountability and reproducibility and thereby helping researchers and practitioners to select appropriate tools and datasets to mitigate these[48]. Functional benefits thus include better AI systems that are better tailored to their users' needs[54], that are more responsible[44] and thus perceived to be legitimate[55]. The final set of 37 documents that fulfilled our criteria of representing AI-IAs turned out to be highly heterogenous. They included short blog posts as well as elaborate documents. Many were presented as separate files in pdf formats, but some were websites, online surveys or spreadsheets containing evaluation criteria. Some had undergone peer review and were published in academic journals, but most were published on the websites of the organisations that had compiled them. We found IA-Ais originating from academic institutions (XXX), public bodies (XXX), standardisation and professional bodies (XXX), civil society organisations (XXX) and companies (xxx). However, these boundaries are not clearly drawn with authorship and ownership of the documents often transcending boundaries. The heterogenous nature of the documents furthermore meant that the application of inclusion and exclusion criteria in many cases required deliberation that led to individual judgement calls.

The functional benefits of AI-IAs can easily be translated into benefits for organisations using AI. Making use of AI-IAs is portrayed as a way of improving organisational processes[35] that support reflection[44] and awareness raising[46,56] and help identify concerns. The use of assessments promises to strengthen robust governance structures[54] that promote organisational oversight[55], help the organisation define its ethical framework[44] and ensure compliance with current as well as future regulation[53]. Having these mechanisms in place is described as a source of competitive advantage for private companies[56,57] and good practice in the public sector[58].

In addition to benefits for organisations, the AI-IAs analysed list benefits for individuals and society. Individuals can benefit by strengthening their rights as data subjects[56] and safeguarding their dignity and human rights[19,55] and their wellbeing[59]. These individual benefits scale on a societal level to the support of fundamental rights more generally[11,28,47]. In addition, societal benefits can include the promotion of particular policy goals that can range from furthering the Sustainable Development Goals[44,56,60] to the more immediate vicinity of AI policy that covers the promotion of responsible innovation[44], increase in trust and avoidance of backlash against new technologies[35].

There are different views of what constitutes or is conceptualised as an AI-IA. They are frequently described as tools[56], which often take the form of self-assessments[19] that can be used for various purposes, such as audits[53] and meeting legal or other requirements (e.g., standards). The description of many AI-IAs makes significant use of the concept of risk management[44,53,61]. AI-IAs are described as facilitating risk estimation[62], risk analysis[48], audit[48] and mitigation[52,53].

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Scope

The AI-IAs define their scope in different ways. Most of them include reference to the technology covered, the application area or domain or the uses of technology. In many cases they cover more than one of these. In some cases, this is done as an explicit delimitation of the scope of the document, whereas others explain the scope through examples or case studies.

The technical scope described in the AI-IAs, not surprisingly, has an emphasis on AI[11,60]. It is worth noting, however, that the terminology is not used uniformly with some documents using terms such as 'intelligent systems'[24,59], 'algorithmic systems'[63] or 'automated decision systems'[35]. In some cases particular types of AI are referred to, notably 'machine learning'[47,54] or relevant features of AI, such as the ability to learn[60] or autonomy[52]. While this focus is dominant, there are references to broader families of technology, such as emerging[52] or disruptive[64] technologies. We also found references pointing beyond particular technologies to the technology ecosystem in which AI is used[65].

The second group of delimitations of scope refers to the application area or domain where the AI is to be applied. It is a frequent occurrence for an AI-IA document to highlight the importance of the domain and list a number of possible domains calling for particular attention[35,36,44,50–52,54,59,65,66]. Among the domains explicitly named, one can find many of those discussed in the media, such as healthcare[19,56,59,64,66–68], finance [50,67,69], security and law enforcement[19,55,61,64], but also other domains, such as education[59,67,68], transport[50,52] and public services[35,59,63,70].

A final set of delimitations of the scope points to specific uses of AI that are deemed to be problematic and in need of an AI-IA[66,68]. These include highly contested uses of AI, for example for surveillance using facial recognition[64], natural language processing[66] or cybersecurity[50].

Issues

The AI-IAs cover a broad range of issues, which can be grouped into the following categories: human rights, ethics, data protection and privacy, security, safety, and environmental impacts. The most frequent topic explicitly referenced is human (or fundamental) rights[11,19,22,35,44,50,51,53,55,56,58,59,61,64,66–69,71–73], with numerous citations to rights as articulated in core international human rights documents (e.g., International Covenant on Economic and Social Rights) and the EU Charter of Fundamental Rights. When assessing ethics[11,19,22,24,35,44,54,58,60,61,64,65,67,68,70,73,74], the most common ethical issues referenced are bias and non-discrimination, fairness and misuse of personal data. Closely related are issues related to data protection and privacy[19,35,44,51,53–58,61,64,65,67], with about half the AI-IA referencing legal compliance obligations, most frequently those under the EU General Data Protection Regulation (GDPR). Fewer AI-IA include dedicated discussion on safety[44,50,54,57,62,65,74] or security[35,44,50,53,54,56,57,62,66], the former focused on harm to human resulting from AI systems and the latter concerned with vulnerabilities of the AI system itself. The final category of issues – environmental impacts[44,58,59] – was less frequently included. Additional issues outside of these categories, mentioned only once or twice, include impacts on the labour market and employment[44,54], accuracy of AI systems[44], and impacts on Western democratic systems[65].

Organisational Context

AI-IAs can be embedded in organisational processes and structures in various ways. They can be viewed as part of a broader governance system[53,55] that contributes to AI's responsible governance[54]. AN AI-IA might be embedded in existing processes, including design, assessment, and marketing of an AI system[56], quality assurance [48], or any existing pre-acquisition assessment[35]. But IA-IAs can also be used on their own[56]. An AI-IA is sometimes carried out by a dedicated team from within the organisation[48] or an external body[19], or both, in those cases where the AI IA includes a self-assessment phase and an assessment by other stakeholders[35]. If the AI-IA is an internal process, the documents we reviewed note the risks of a conflict of interest or a lack of independence of the body implementing it[19,65].

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

The responsibility for the AI is described as falling on the organisations using it, and they are the ones responsible for the IA[35,58,60]. The documents we reviewed suggest that public bodies should be required to conduct self-assessment of AI systems[11,35]. At the same time, different aspects of responsibility for ensuring that AI-IA is completed reside with various actors. For example, governments are responsible for setting out procedures for public authorities to carry out an assessment[11]and due mechanisms for affected individuals or communities to participate in it[35].

AI-IAs have roots in the tradition of impact assessments in different domains, particularly environmental protection, human rights, and privacy[35,55]. Further IAs that the AI-IAs can draw from, overlap and sometimes complements can be broadly grouped into two categories. The first are IAs mainly interested in data: data protection impact assessments (DPIA)[19,24,53,71], privacy impacts assessments (PIA)[19,55] or surveillance impact assessment[55]. The second category are IAs that focus on societal and ethical impacts. These include ethical impact assessments[19,55], societal impact assessments[19], and equality impact assessments[58]. The assessments differ in terms of their mandatory or voluntary nature[19]. It has been suggested that AI-IAs may be integrated with the DPIA[53]. In contrast to DPIAs, AI-IA are rarely mandatory[71]. What distinguishes AI-IAs from other impact assessments is the fact that they are technology-specific.

Timeframe

Regarding the timing of potential AI impacts, only one AI-IAs recognized the need to distinguish between short, medium, and long-term risks[65]. In terms of the point at which the IA is carried out, if the AI is purchased from another organization, it has been suggested that the IA is implemented before the AI deployment[35,73] or, when possible, before its acquisition[11,35]. In the case of organizations that design and develop AI, the IA is recommended at the beginning of the project[36,60]. Besides the start of a project, the documents we analysed suggest the AI-IA is carried out regularly at several other points of the AI lifecycle[11,35]. It is advised that AI-IA is revisited and revised at each new phase of AI lifecycle[11], when significant changes are introduced[52], e.g., changes to data collection, storage, analysis or sharing processes[58] and before the production of the system[36]. It has been suggested that the assessment be renewed at a set time, every couple of years[35]. There seems to be a consensus that AI-IA should be iterative, and the new iterations should be informed by contemporary research and feedback from the AI implementation[44,59,71].

Process and methods

Having a recognisable process that allows users to undertake an AI-IA was a criterion for including a document in our analysis which ensured they all provided some practical guidance. The structure and detail of the processes covered differ greatly. Most of the IA-IAs describe an explicit structure consisting of phases or steps associated with an AI-IA[24,35,52,59,60,63]. The can start with the determination of what counts as acceptable uses of AI[64] which can draw upon shared values and principles[19]. This can be part of the preparatory activities of an AI-IA which can also include a definition of benefits expected from the AI[58] and the need for the impact assessment[60] as well as the development of skills required to undertake it[53]. A further preliminary step is the attribution of responsibility for the AI-IA[54,57,60].

The practical steps can start by setting up procedures for documentation and accountability[60] as well as a description of the AI in question [55,60] and the justification of its use[60]. A core component of the AI-IAs is a set of questions in the form of a questionnaire or checklist that the AI-IA seeks responses to[19,48,51,55]. These questions are often justified on the basis of existing normative guidance ranging from human rights [44,56,69] and existing legislation such as the GDPR[55] to lists of ethical issues[62], principles of sustainability[35,44] and responsible innovation[48]. These questions cover the various issues associated with AI such as data protection[35,44,55], data quality and representativeness of data[58], fairness[69], reproducibility[58], explainability[58], transparency and accessibility[44], often recognising that there are trade-offs between some of these issues[53]. Often these questions lead to a quantitative scoring of issues and risks[36,53] or the determination of key

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

performance indicators. These draw on scientific insights[35,44,59] from various disciplines, such as psychology[24] or foresight analysis[52].

A further aspect that is shared by many of the AI-IAs is the inclusion of stakeholders in the assessment process[19,35,44,48,55,59]. Considerable effort is spent on the identification of suitable stakeholders who are typically expected to cover the relevant areas of expertise of the AI application as well as the groups affected by it. Examples of such stakeholder groups include AI users[66], external experts[55,63], technology providers[60], senior manager[53] and civil society more broadly[55].

Following the identification of issues, most AI-IAs proceed to outline specific steps that can be used to mitigate undesirable consequences of AI[11,22,52,53]. There are numerous categories of mitigation measures[55,60,66] including technical measures such as de-biasing training data[22] or code inspections[63] and organisational measures[44,57] such as the creation of accountability structures[44], documentation[58], evaluation and monitoring of systems use[58] but also enabling human interventions[22]. One can find suggestions to inclusion and diversity[44], promote training and education of the workorce[52,58,66], the inclusion of external experts[55] or the definition of redress mechanisms[44]. These mitigation measures all suffer, however, from the uncertainty of future occurrences[64] which can require situation-specific responses[64] and call for the maintenance of mitigation mechanisms over time[65].

Transparency

The AI-IA documents share a common standpoint over the importance of transparency and communication in AI systems. Transparency means that actions, processes and data are made open to inspection by publishing information about the project in a complete, open, understandable, easily-accessible, and free format[58].

The key is to help humans understand why a particular decision has been made, and provide the confidence that the AI model system has been tested and makes sense. Transparency about how an AI application works gives individuals the opportunity to appreciate the effects of the application on the freedom of action and the room to make decisions[60]. In practice, this can mean various things. It may mean that there is access to the source code of an AI application, that to a certain extent, end-users are involved in the design process of the application, or that an explanation is provided in general terms about the operation and the context of the AI application. Transparency about the use of AI applications may enlarge the individual's autonomy, because it gives the individual the opportunity to relate to, for instance, an automatically made decision[60].

However, limitations in the ability to interpret AI decisions is not only frustrating for end-users or customers, but can also expose an organisation to operational, reputational, and financial risks[54]. To instil trust in AI systems, people must be enabled to look “under the hood” at their underlying models, explore the data used to train them, expose the reasoning behind each decision, and provide coherent explanations to all stakeholders in a timely manner[54]. Individuals must perceive that they have a reasonable voice in the decision-making process, that the decision-makers have treated them respectfully, and the procedure is one they regard as fair[64].

A trustworthy approach is key to enabling ‘responsible competitiveness’, by providing the foundation upon which all those using or affected by AI systems can trust that their design, development and use are lawful, ethical and robust[44]. A crucial component of achieving Trustworthy AI is transparency which encompasses three elements: 1) traceability, 2) explainability and 3) open communication about the limitations of the AI system[44].

The starting point of the Impact Assessment is that with every deployment of AI, the organisation looks at what is required for transparency and what that means for the design of the technique, the organisation or

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

the people working with the technology[60]. For example, companies must publicly disclose information about each automated decision system, including details about its purpose, reach, potential internal use policies or practices, and implementation timeline[35]. The initial disclosure provides a strong foundation for building public trust through appropriate levels of transparency, while subsequent requests can solicit further information or the presentation of new evidence, research, or other inputs that the agency may not have adequately considered otherwise[35].

Currently, few agencies are explicitly mandated to disclose anything about the systems they have in place or are planning to use. Instead, impacted communities, the public at large, and governments are left to rely on what journalists, researchers, and public records requests have been able to expose[35]. However, government bodies and external auditors can play a crucial role in enabling open transparency between the AI technology and its users, but it is important that robust processes are in place to carry out the audit effectively. This means auditing tools must be explicit and clear about which definitions they evaluate, what those definitions mean, and in what ways they are limited[73]. Auditing must fit within a broader approach to evaluating the impact of AI systems on equality. This comprehensive evaluation should include reasonable consideration of impacts on equality of opportunity and outcome, and focus companies on the making of adjustments to mitigate relevant adverse impacts which have been identified[73]. Furthermore, the auditors must live up to an ethical standard themselves in order to enhance fairness and evaluate the impact of the AI system over time.

Challenges

Assessing the impact of AI raises significant challenges, starting from the variety of AI applications themselves, which makes it more difficult to understand the nature of AI and its consequences and how these are reflected in social norms[64]. For example, assessing the impact of an AI solution may involve the consideration of fairness in terms of the existence of bias, but it may involve trade-offs that render it impossible to be fair to everybody[54]. However, even though there is continuous demand for more greater regulation[24], the arguments on the flipside, e.g. that such regulation slows innovation agility are increasing. The open nature of AI as a general purpose technology renders prediction of consequences difficult, which contributes to challenges of governance [48]

Assessing AI impact, considering both ethics and innovation is an important part of an AI impact assessment, but the impact itself is difficult to model[24], especially because AI-based systems are not static, as usually assumed by traditional impact assessments; instead they are very dynamic as they are adding new data, learning and refining models[24]. In addition, to accurately capturing the system itself, attention must be given to the way that the system is used in a particular organisation and the structure of any impact assessment procedure such that it does not end up being excessively burdensome and complex[19]. Additionally, defining values as benchmarks in an impact assessment procedure becomes challenging just because of the variety and complexity of such values, and the need to tailor them to the specific application[19]. This refined assessment approach may generate additional burden to companies as they may be expected to broadly identify and mitigate every conceivable kind of risk[66].

Discussion

Our analysis has shown that there is broad interest in AI-IAs from various quarters. AI-IAs offer a practical approach to the ethical and social issues of AI that is missing from the guideline-centric approach that currently dominate the debate[25]. Our research suggests that there is a certain level of convergence between AI-IAs. However, the research also shows that a number of open questions remain.

A first set of questions pertain to concepts and definitions. While AI is broadly discussed and definitions of AI abound, there is no universally accepted and unambiguous definition of AI[75] which renders it difficult to delineate the exact scope of an AI-IA. This is reflected in the titles of many of the documents we reviewed,

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

which use other terms like 'algorithm' or 'big data'. These other terms do not solve the problem, as they introduce new types of ambiguity. Exact definitions of terms are usually difficult to agree on. In the case of AI-IAs this lack of a clear definition of the technology that it refers to is problematic for several reasons. On the one hand a broad definition of the underlying technology may call for a sweeping application of such AI-IAs which could be prohibitively costly and at the same time not plausible. If, for example one were to undertake a full impact assessment of all technical systems that are based on or incorporate algorithms, then this would cover most outputs of computer programming which would be far too broad. A narrow scope, for example one focusing on particular types of applications of deep learning only, might miss new developments and therefore not capture developments that have significant potential for risk. A further problem of the lack of a clear definition of AI is that it renders a general application of AI-IAs unlikely, as owners and users of AI may justifiably argue that it is not clear which systems exactly are to be subject of such an assessment.

Further conceptual questions arise with regards to the scope and scale of AI-IAs. Some of the document we analysed have a broad scope and ambition whereas others focus on specific applications or issues. Some are predominantly focused on the technology in question whereas others think more broadly in terms of organisational embedding of technology, required capacities by staff to deal with them etc. This breadth of scope is not problematic per se, but it raises the question how many AI-IAs are needed. A large number may be useful in catering for many applications, but it has the disadvantage of making it difficult for potential users to understand the landscape.

A further fundamental question is whether a particular AI will have an impact at all or an impact that calls for an AI-IA. Any use of an AI is of course expected to have some impact; otherwise, there would be no point in employing it. However, only when there is reason to believe that an AI is likely to lead to socially or ethically relevant change does it make sense to consider whether these changes are positive, negative, call for mitigation measures etc. Impact, in many cases, can be defined rigorously, though what definitions optimally capture the most important aspects in a given use case can be a challenging question. A good example of impact definition is provided by Berk[76], in the context of the use of machine learning forecasts by a parole board to help inform parole release decisions. The paper defines and evaluates the impact of the forecasts through stating and addressing the following three questions: Did the overall proportions of inmates released by the Board change because of the forecasts? Did the forecasts lead to changes in the kinds of inmates the Board released on parole? What impact, if any, did the forecasts have on arrests after an individuals were paroled?

Defining impact in such a manner can enable us, in principle, to evaluate it via statistical hypothesis testing. A key challenge in applying a mathematically rigorous method is, of course, the availability of datasets satisfying certain requirements. In the case described by Berk[76], for example, because the machine learning system was introduced into the Board operations gradually, it was possible to split a large set of parole cases into the treatment group and the comparison group, and the randomness assumption about the composition of the groups appeared plausible. While such datasets may not always be readily available for deployed AI-powered systems, we think that their designers, integrators and operators often have sufficient control for enabling more rigorous system's impact assessment.

The question whether an AI has an impact introduces numerous additional considerations. One observation from our analysis is that many of AI systems under discussion are still under development or found in a research setting. In such cases even the intended outcomes may not be clear which makes it difficult to determine which impacts to look for. The aim of AI-IAs on delivering technical, individual, organisational but also societal benefits makes the determination of relevant impacts difficult. Many of the documents we analysed refer to ethical principles or human rights. In some cases, the impacts on these will be possible to

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

capture, as the earlier example of parole decisions indicates. In other cases where impacts are on broader concepts, such as human dignity or societal justice, this will be more difficult.

The topic of measuring impacts leads to questions of trade-offs within AI-IAs as well as the cost-benefit balance of the AI-IA approach as a whole. Trade-offs can be expected in many impact assessments where an aspect deemed desirable leads to consequences that are undesirable. In AI, for example, it is likely that trade-offs will appear between privacy of individuals versus transparency of the AI. Many similar trade-offs are conceivable and should be captured and evaluated by an AI-IA. The cost-benefit balance of the AI-IA approach as a whole is a special type of trade-off. The benefits of an AI-IA not only depend on the identifiability of impacts but also on whether the impact assessment has consequences that support desired impacts. Measuring such impacts will be difficult if not impossible. This is caused by the potential of long-term impact which is difficult to measure in the short term and may be impossible to measure at all or to quantify. The costs of undertaking an AI-IA may be easy to measure on an organisational level. However, in addition to the immediate financial costs of undertaking an AI-IA, there may be side effects, such as a slowing down of the rate of innovation or the self-censoring of innovators which can be counted as further costs on a societal level that may also be impossible to measure.

Such questions are of course not confined to AI-IAs, but similarly apply to other types of impact assessment or risk management measures. It is therefore important to consider the embedding of AI-IAs in existing structures. Our analysis has shown that many AI-IAs reference other types of impact assessment and it therefore seems reasonable to embed them in established activities, such as due diligence or risk management processes which may already cover environmental or other impact assessments. One important part of the discussion that has the potential to significantly affect the cost-benefit analysis from an organisational point of view is that of sanctions for undertaking (or omitting) AI-IAs. If an organisation could be fined or if its liability threshold were to change because of an AI-IA, this would change its willingness to undertake one. Interestingly, however, our analysis of the existing AI-IAs found very little reference to such external sanctions. The majority of the AI-IAs we investigated relied on positive messages and the benefits of AI-IAs with little reference to legal or other mandates to undertake them or negative sanctions for failing to do so.

The current landscape of AI-IAs thus retains numerous open questions. While significant efforts have been undertaken in defining and trialling such IAs, there remain a number of concerns. Existing AI-IAs are intended to do good, but it is often not clear who will benefit from them or how competing interests are considered, e.g. when organisational benefits conflict with societal ones. The current landscape furthermore shows the danger of fragmentation. Our sample includes 37 AI-IAs and we can expect the number to grow. This leads to problems of choosing an appropriate AI-IA for the user. Maybe more importantly, it makes it difficult to assess who will benefit from applying any individual AI-IA. In addition, the application of AI-IAs is fraught with uncertainty and subjectivity. Many of the aspects of AI-IAs are open to interpretation. Abstract criteria and grading scales are sometimes provided but grading can be highly subjective. There is a trade-off to pay for being generic and proposing an IA process that can be applied to virtually any use-case and scientific precision which may be impossible to achieve.

These concerns lead to a larger one that AI-IAs will be used for what is sometimes called 'ethics washing'[77]. It has been observed by several authors that the AI ethics debate is in constant danger of being hijacked by particular interests, in particular the interests of large corporation who have a vested interest in using ethical rhetoric to avoid regulation and deflect scrutiny[7,78–80]. The use of AI-IAs would be a good tool for such purposes, as it remains within the remit of the organisation doing it to undertake it and to disseminate it. As we have shown, there is a strong emphasis on transparency of findings and broad stakeholder inclusion in many of the AI-IA processes we have investigated, both of which can be read as mechanisms to avoid the dominance of vested interests. It is not clear, however, whether they will suffice or whether

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

independent and maybe governmental control, regulation and oversight would be required to address this concern.

A final concern worth highlighting is that of the functional or techno-optimist underpinnings of AI-IAs. The majority of the documents we investigated started by outlining the benefits of AI, balances these against the downsides and then suggests that an AI-IA is a mechanism that will increase the likelihood that the benefits can be retained while managing risks and downsides. This is the techno-optimist view that AI is fundamentally an ethically and socially good thing. In this mindset AI-IAs are purely functional tools to ensure that AI's benefits can unfold. This narrative pervades the AI literature and in particular the AI policy landscape. It is, however, by no means certain that this is the only or best framing of AI in general or of specific AI technologies and applications. It may very well be that the world would be better off without AI or some particular technologies. This is of course a much broader societal debate, but AI-IAs, by offering a tool to address the downsides of AI, may stifle this broader societal debate about what future we are collectively trying to achieve, and which role technology should play in that future.

Conclusion

This paper offers the first systematic review of AI-IAs. In light of growing interest in not only the ethics of AI but also regulation of AI, it can be expected that AI-IAs are likely to play an important role in future AI governance. The paper therefore will be of interest to researchers working on AI ethics, and AI policy. It also makes a practical contribution that is relevant to both policymakers who are considering how to implement AI policies and organisations interested in using an AI-IA to better understand and reflect on their technologies or aiming to broaden their risk management processes.

As any other research, this paper has limitations. We set out to undertake a systematic review of AI-IAs. However, the nature of these documents renders it difficult to arrive at an incontrovertible population of documents. We believe that our multi-pronged search strategy allowed us to identify all, or at least the most relevant AI-IAs. However, we cannot prove this and new AI-IAs may have become available since we undertook the search in the European spring of 2021. In addition, the conceptual fuzziness of AI means that it is very difficult to precisely delineate the inclusion and exclusion criteria. Due to our search strategies, our sample ended up including some documents that focus on closely related questions such as data ethics[58] and were found to fall within our definition of AI-IAs, but we concede that different interpretations would be possible, leading to a different population of AI-IAs. It is unlikely, however, that the inclusion of additional AI-IAs or the removal of parts of the documents we analysed would fundamentally alter our findings.

This paper should provide a sound basis for the next step in developing AI-IAs. The documents we have analysed include several well-researched, mature, and reflected examples which can be implemented by organisations. What seems to be missing at the moment is a more comprehensive overview of their role in the AI ecosystem. We have shown that there is much attention to other types of impact assessments, calls for the coordination with such impact assessment, consideration of the integration of AI-IAs into other organisational processes such as risk management, as well as numerous references to relevant regulation. It is thus clear that AI-IAs need to be understood in this broader context.

At present, however, there is very little guidance on the role of AI-IAs in the broader context of the AI innovation ecosystems. This makes it difficult for organisations planning to use AI to identify the most appropriate AI-IA for their specific needs. This contributes to the challenge of evaluating whether a particular AI-IA is fit for purpose and whether an organisational application of it can or will have the desired outcome.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Some of these problems are likely to be temporary and upcoming legislation, regulation, professional guidance and case law will make the role of AI-IAs in their ecosystems clearer. At the same time there is need for research to better understand the impact of AI-IAs. They are typically framed in terms of the benefits they offer for individuals, organisations and society as a whole. What is currently unclear is whether the application of an AI-IA actually leads to the promised benefits and how this could be measured. Such research is urgently needed to ensure that AI-IAs can contribute to addressing the ethical and social consequences of AI use, while simultaneously not overloading them with unachievable expectations. We hope that this research has provide a robust evidence base for such further research and thereby contributes to the overall aim of ensuring that AI contributes to human flourishing.

References

1. Access Now Policy Team. *The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems*. https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf (2018).
2. Stone, P. *et al.* Artificial Intelligence and Life in 2030. One hundred year study on artificial intelligence: Report of the 2015-2016 Study Panel. *Stanford University, Stanford, CA, http://ai100. stanford. edu/2016-report. Accessed: September* **6**, 2016 (2016).
3. Persson, A. Implicit Bias in Predictive Data Profiling Within Recruitments. in *Privacy and Identity Management. Facing up to Next Steps* (eds. Lehmann, A., Whitehouse, D., Fischer-Hübner, S., Fritsch, L. & Raab, C.) 212–230 (Springer International Publishing, 2016). doi:10.1007/978-3-319-55783-0_15.
4. Muller, C. *The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law*. https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da (2020).
5. Zuboff, P. S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. (Profile Books, 2019).
6. Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat Mach Intell* **1**, 389–399 (2019).
7. Mittelstadt, B. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* **2019**, (2019).
8. Clarke, R. Principles and Business Processes for Responsible AI. *Computer Law & Security Review* **35**, 410–422 (2019).

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

9. Clarke, R. Regulatory Alternatives for AI. *Computer Law & Security Review* **35**, 398–409 (2019).

10. Access Now. *Human Rights in the Age of Artificial Intelligence*. https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf (2018).

11. Council of Europe. *Unboxing artificial intelligence: 10 steps to protect human rights*. https://www.coe.int/en/web/commissioner/view/-/asset_publisher/ugj3i6qSEkhZ/content/unboxing-artificial-intelligence-10-steps-to-protect-human-rights (2019).

12. Erdélyi, O. J. & Goldsmith, J. Regulating Artificial Intelligence: Proposal for a Global Solution. in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* 95–101 (Association for Computing Machinery, 2018). doi:10.1145/3278721.3278731.

13. Wallach, W. & Marchant, G. Toward the Agile and Comprehensive International Governance of AI and Robotics [point of view]. *Proceedings of the IEEE* **107**, 505–508 (2019).

14. UNESCO. *First version of a draft text of a recommendation on the Ethics of Artificial Intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000373434 (2020).

15. Expert Group on Liability and New Technologies. *Liability for Artificial Intelligence and other emerging digital technologies*. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=63199 (2019).

16. OECD. *Recommendation of the Council on Artificial Intelligence*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (2019).

17. Becker, H. A. Social impact assessment. *European Journal of Operational Research* **128**, 311–321 (2001).

18. Hartley, N. & Wood, C. Public participation in environmental impact assessment—implementing the Aarhus Convention. *Environmental Impact Assessment Review* **25**, 319–340 (2005).

19. Mantelero, A. AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review* **34**, 754–772 (2018).

20. CNIL. *Privacy Impact Assessment (PIA) Methodology*. http://www.cnil.fr/fileadmin/documents/en/CNIL-PIA-1-Methodology.pdf (2015).

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

21. Information Commissioner's Office. *Privacy Impact Assessment Handbook, v. 2.0*. http://www.ico.gov.uk/upload/documents/pia_handbook_html_v2/files/PIAhandbookV2.pdf (2009).

22. Ivanova, Y. The Data Protection Impact Assessment as a Tool to Enforce Non-discriminatory AI. (2020).

23. CEN-CENELEC. *Ethics assessment for research and innovation - Part 2: Ethical impact assessment framework*. ftp://ftp.cencenelec.eu/EN/ResearchInnovation/CWA/CWA17214502.pdf (2017).

24. Calvo, R. A., Peters, D. & Cave, S. Advancing impact assessment for intelligent systems. *Nature Machine Intelligence* **2**, 89–91 (2020).

25. Stix, C. Actionable Principles for Artificial Intelligence Policy: Three Pathways. *Sci Eng Ethics* **27**, 15 (2021).

26. UK AI Council. *AI Roadmap*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/949539/AI_Council_AI_Roadmap.pdf (2021).

27. EDPS. *A Preliminary Opinion on data protection and scientific research*. https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf (2020).

28. FRA. *Getting the future right – Artificial intelligence and fundamental rights*. https://fra.europa.eu/en/publication/2020/artificial-intelligence-and-fundamental-rights (2020).

29. European Commission. *Proposal for a Regulation on a European approach for Artificial Intelligence*. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence (2021).

30. Boell, S. K. & Cecez-Kecmanovic, D. On being 'systematic' in literature reviews in IS. *J Inf technol* **30**, 161–173 (2015).

31. Rowe, F. What literature review is not: diversity, boundaries and recommendations. *European Journal of Information Systems* **23**, 241–255 (2014).

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

32. Schwarz, A., Mehta, M., Johnson, N. & Chin, W. W. Understanding frameworks and reviews: a commentary to assist us in moving our field forward by analyzing our past. *SIGMIS Database* **38**, 29–50 (2007).

33. Stahl, B. C., Timmermans, J. & Mittelstadt, B. D. The Ethics of Computing: A Survey of the Computing-Oriented Literature. *ACM Comput. Surv.* **48**, 55:1-55:38 (2016).

34. Liberati, A. *et al.* The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLOS Medicine* **6**, e1000100 (2009).

35. AI Now Institute. *ALGORITHMIC IMPACT ASSESSMENTS: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY ACCOUNTABILITY*. https://ainowinstitute.org/aiareport2018.pdf (2018).

36. Corriveau, N. *The Government of Canada's Algorithmic Impact Assessment: Towards Safer and More Responsible AI*. https://aiforsocialgood.github.io/2018/pdfs/track2/83_aisg_neurips2018.pdf (2018).

37. Metcali, J., Moss, E., Watkins, E. A., Singh, R. & Elish, M. C. Algorithmic Impact Assessments and Accountability: *ACM* (2021) doi:https://dl.acm.org/doi/proceedings/10.1145/3442188.

38. Liu, W., Zhao, J., Du, L., Padwal, H. H. & Vadivel, T. Intelligent comprehensive evaluation system using artificial intelligence for environmental evaluation. *Environmental Impact Assessment Review* **86**, (2021).

39. Park, D. & Um, M.-J. Robust Decision-Making Technique for Strategic Environment Assessment with Deficient Information. *Water Resources Management* **32**, 4953–4970 (2018).

40. IEEE Computer Society. *IEEE Standard Model Process for Addressing Ethical Concerns during System Design - 7000-2021*. https://engagestandards.ieee.org/ieee-7000-2021-for-systems-design-ethical-concerns.html (2021).

41. IAIA. Impact Assessment. https://www.iaia.org/wiki-details.php?ID=4.

42. Aronson, J. A pragmatic view of thematic analysis. *The qualitative report* **2**, 1–3 (1995).

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

43. Braun, V. & Clarke, V. Using thematic analysis in psychology. *Qualitative research in psychology* **3**, 77–101 (2006).

44. AI HLEG. *Assessment List for Trustworthy AI (ALTAI)*. https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence (2020).

45. QSR. NVivo 11 for Windows Help - Run a coding comparison query. http://help-nv11.qsrinternational.com/desktop/procedures/run_a_coding_comparison_query.htm.

46. UnBias. Fairness Toolkit. *UnBias* https://unbias.wp.horizon.ac.uk/fairness-toolkit/ (2018).

47. Winter, P. *et al. White Paper - Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications*. 47 https://www.tuv.at/loesungen/digital-services/trusted-ai/ (2021).

48. Raji, I. D. *et al.* Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 33–44 (ACM, 2020). doi:https://doi.org/10.1145/3351095.3372873.

49. IEEE. *IEEE 7010-2020 - IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being*. https://standards.ieee.org/content/ieee-standards/en/standard/7010-2020.html (2020).

50. Government Accountability Office. *Technology Assessment, Emerging Opportunities, Challenges, and Implications*. (2018).

51. Gebru, T. *et al.* Datasheets for Datasets. *arXiv:1803.09010 [cs]* (2020).

52. Brey, P. D6.1: Generalised methodology for ethical assessment of emerging technologies. (2020).

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

54. PWC. *A practical guide to Responsible Artificial Intelligence (AI)*. https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai/responsible-ai-practical-guide.pdf (2019).

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

55. Kaminski, M. E. & Malgieri, G. Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations. (2019) doi:10.2139/ssrn.3456224.

56. Williams, carmel. A Health Rights Impact Assessment Guide for Artificial Intelligence Projects. - Abstract - Europe PMC. *Health and Human Rights Journal* **22**, 55–62 (2020).

57. PricewaterhouseCoopers. Responsible AI Toolkit. *PwC* https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html (2019).

58. UK Governmental Digital Service. *Data Ethics Framework*. 38 (2020).

59. IEEE. *IEEE 7010-2020 - IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being*. https://standards.ieee.org/content/ieee-standards/en/standard/7010-2020.html (2020).

60. ECP Platform for the Information Provision. *Artificial Intelligence Impact Assessment*. https://ecp.nl/wp-content/uploads/2019/01/Artificial-Intelligence-Impact-Assessment-English.pdf (2019).

61. Oswald, M. Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality. (2018) doi:https://doi.org/10.1080/13600834.2018.1458455.

62. Devitt, K., Gan, M., Scholz, J. & Bolia, R. *A Method for Ethical AI in Defence*. 76 https://www.dst.defence.gov.au/publication/ethical-ai (2020).

63. Ada Lovelace Institute. *Examining Tools for assessing algorithmic systems the Black Box*. https://www.adalovelaceinstitute.org/wp-content/uploads/2020/04/Ada-Lovelace-Institute-DataKind-UK-Examining-the-Black-Box-Report-2020.pdf (2020).

64. Deloitte Australia. *A moral license for AI - Ethics as a dialogue between firms and communities*. https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/artificial-intelligence-impact-on-society.html (2020).

65. Zicari, R. V. *et al.* Z-Inspection®: A Process to Assess Trustworthy AI. *IEEE Transactions on Technology and Society* **2**, 83–97 (2021).

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

66. Andrade, N. N. G. & Kontschieder, V. *AI Impact Assessment: A Policy Prototyping Experiment*. https://openloop.org/wp-content/uploads/2021/01/AI_Impact_Assessment_A_Policy_Prototyping_Experiment.pdf (2021).

67. Raso, F. A., Hilligoss, H., Krishnamurthy, V., Bavitz, C. & Kim, L. *Artificial Intelligence & Human Rights: Opportunities & Risks*. https://papers.ssrn.com/abstract=3259344 (2018) doi:10.2139/ssrn.3259344.

68. Gardner, A., Smith, A. L., Steventon, A., Coughlan, E. & Oldfield, M. Ethical funding for trustworthy AI: proposals to address the responsibilities of funders to ensure that projects adhere to trustworthy AI practice. *AI Ethics* (2021) doi:10.1007/s43681-021-00069-w.

69. Schmitt, C. E., September 27, & 2018. Evaluating the impact of artificial intelligence on human rights. (2018).

70. Leslie, D. Understanding artificial intelligence ethics and safety. (2019).

71. Europäische Union & Agentur für Grundrechte. *Getting the future right artificial intelligence and fundamental rights ; report*. (2020).

72. Microsoft and Article One. Human Rights Impact Assessment (HRIA) of the Human Rights Risks And Opportunities Related To Artificial Intelligence (AI). (2018).

73. Institute for the future of work. *Artificial intelligence in hiring - assessimg impacts on equality*. https://www.ifow.org/case-studies/introducing-equality-impact-assessments-for-artificial-intelligence (2020).

74. AI Now Institute. Algorithmic Impact Assessments: Toward Accountable Automation in Public Agencies. https://medium.com/@AINowInstitute/algorithmic-impact-assessments-toward-accountable-automation-in-public-agencies-bd9856e6fdde (2018).

75. Elsevier. *Artificial Intelligence: How knowledge is created, transferred, and used - Trends in China, Europe, and the United States*. https://www.elsevier.com/?a=827872 (2018).

76. Berk, R. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *J Exp Criminol* **13**, 193–216 (2017).

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

77. Wagner, B. Ethics as an escape from regulation: From ethics-washing to ethics-shopping. in *Being Profiled: Cogitas Ergo Sum* (eds. Bayamlioglu, E., Baraliuc, I., Janssens, L. A. W. & Hildebrandt, M.) 84–90 (Amsterdam University Press, 2018).

78. Nemitz, P. Constitutional democracy and technology in the age of artificial intelligence. *Phil. Trans. R. Soc. A* **376**, 20180089 (2018).

79. Findlay, M. & Seah, J. An Ecosystem Approach to Ethical AI and Data Use: Experimental Reflections. in *2020 IEEE / ITU International Conference on Artificial Intelligence for Good (AI4G)* 192–197 (2020). doi:10.1109/AI4G50087.2020.9311069.

80. Coeckelbergh, M. Artificial Intelligence: Some ethical issues and regulatory challenges. *1* 31–34 (2019) doi:10.26116/techreg.2019.003.

81. Stahl, B. C. *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*. (Springer International Publishing, 2021). doi:10.1007/978-3-030-69978-9.

82. Stahl, B. C. *et al.* Artificial intelligence for human flourishing – Beyond principles for machine learning. *Journal of Business Research* **124**, 374–388 (2021).

83. AIEI Group. *From Principles to Practice - An Interdisciplinary framework to operationalise AI ethics*. 56 https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf (2020).

84. Brinkman, B. *et al.* Listening to Professional Voices: Draft 2 of the ACM Code of Ethics and Professional Conduct. *Commun. ACM* **60**, 105–111 (2017).

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

86. Martin, C. D. & Makoundou, T. T. Taking the high road ethics by design in AI. *ACM Inroads* **8**, 35–37 (2017).

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

87. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. & Srikumar, M. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. (2020).

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641