

Towards accountable algorithms: tools and methods for responsible use

Un programa de



GOBIERNO
DE ESPAÑA

MINISTERIO
PARA LA TRANSFORMACIÓN DIGITAL
Y DE LA FUNCIÓN PÚBLICA

SECRETARÍA DE ESTADO
DE DIGITALIZACIÓN
E INTELIGENCIA ARTIFICIAL

red.es



Mobile
WorldCapital
Barcelona

About Digital Future Society

Digital Future Society is a non-profit transnational initiative that engages policymakers, civic society organisations, academic experts and entrepreneurs from around the world to explore, experiment and explain how technologies can be designed, used and governed in ways that create the conditions for a more inclusive and equitable society.

Our aim is to help policymakers identify, understand and prioritise key challenges and opportunities now and in the next ten years in the areas of public innovation, digital trust and equitable growth.

Visit digitalfuturesociety.com to learn more

A programme of



red.es



Permission to share

This publication is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/) (CC BY-SA 4.0).

Published

March 2024

Disclaimer

The information and views set out in this report do not necessarily reflect the official opinion of Mobile World Capital Foundation. The Foundation does not guarantee the accuracy of the data included in this report. Neither the Foundation nor any person acting on the Foundation's behalf may be held responsible for the use which may be made of the information contained herein.

Table of contents

Six key ideas	5
Introduction	7
1. Why evaluate algorithms in the current context of artificial intelligence?	9
2. Conceptual framework: definitions and dimensions of algorithm evaluations	12
3. Algorithm evaluation methods	21
4. The ecosystem of algorithmic evaluations and governance levels of algorithmic accountability	32
5. Looking to the future: improving algorithmic evaluation processes	42
Conclusions	45
References	46
Appendix	51
Acknowledgements	55

Six key ideas

1. Algorithmic evaluations are a critical issue.

Algorithmic systems have a significant impact on society. They give rise to discrimination and bias, have harmful effects on environmental sustainability and lead to violations of privacy, among other things. All this creates the need to conduct a holistic analysis of algorithmic systems to detect problems and offer risk mitigation measures that strengthen sustainable innovation.

2. Several aspects of algorithmic evaluations need addressing.

Subject to a set of dimensions (focus, locus, stakeholders, timing, topic, scope, etc.), algorithmic evaluations can prioritize different aspects of the process, such as technological issues. But they can also focus on more human attributes that account for the interplay between the technology and the social context in which an algorithm is deployed.

3. There is no single recipe for an algorithmic evaluation.

Available methods vary by approach and come with advantages and limitations in certain contexts. The intersection of methods (code audits, scraping, checklists, case studies, etc.) and the dimensions of algorithmic evaluations, present us with a situational map of opportunities and limitations.

4. Algorithms are not evaluated in a vacuum.

When analysing the functioning and effects of an algorithm, the stakeholder ecosystem that comes into play must be considered. There are three levels of governance: interaction between the public and private sectors and the third sector (macro); activity sectors such as health, education, security, etc. (mezzo); and stakeholders that design, implement, use and audit algorithms (micro). Understanding these dynamics will facilitate more appropriate algorithmic evaluations with fuller accountability.

5. People should always be the focus.

Regardless of the approach followed and the methods used, people must be prioritized. This means we need to aim to understand how algorithms work and the impact they have on lives – especially those of vulnerable and excluded groups. We must also pay attention to the people who design and implement these systems, to their relationships with organizational structures and to the broader social context in which they operate.

6. Algorithm standards and supervisory bodies must be given importance.

The public, private and third sectors must work together to define clear standards in algorithmic evaluation processes. This will prevent inappropriate practices that undermine the processes, while promoting shared criteria to move towards improved algorithmic evaluations. The existence of algorithm supervisory bodies will also contribute decisively to the effectiveness of and confidence in these processes, especially if they promote international cooperation and knowledge exchange.

Introduction

As the implementation of algorithms advances in an increasing number of contexts and sectors, the need for debate on the ethical aspects and governance of artificial intelligence (AI) grows. Amid promises of improved efficiency and effectiveness, algorithmic systems can contain biases and make mistakes that have unwanted effects on people's lives. Algorithmic evaluations can help mitigate these effects by detecting problematic issues such as discrimination against population groups, distortion of reality and exploitation of personal information (Bandy 2021). Specifically, evaluation processes drive compliance with the ethical principles of AI regulations and strategy documents.

In recent years several published academic papers and reports have sought to detail what algorithmic evaluations should include. In practical terms, they aim to answer the following question: **How can algorithms be evaluated to detect any potential problems they may contain and/or issues that may arise from their use, and how can these be mitigated?** Approaches are varied. Some prioritize a mainly technological viewpoint of algorithmic system analysis; others advocate a more general and holistic study of the risks and impacts on populations and organizations. Evaluations can take place before or after the system is implemented, and with or without the participation of external stakeholders.

To bring clarity to this complex issue, this report examines and systematizes existing options for algorithmic evaluations and provides an overview of methods and tools that can be used depending on the evaluator's objectives and available resources. It also explains the ecosystem of stakeholders and sectors involved considering a very general framework for algorithmic accountability. Lastly, it offers six recommendations to improve algorithmic evaluations in the future.

The research for this report followed a three-phase qualitative **methodology** (see Annex):

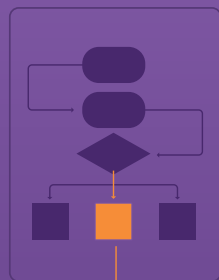
- **Systematic review of the academic literature.**
ASReview, an active learning tool that uses machine learning to select relevant articles, was used to analyse a total of 64 documents from a sample of almost 3,000. The result was a summary of the most recent global research.
- **Documentary analysis of the grey literature.**
We then conducted a search for reports and publications other than academic papers, published and distributed by public bodies, third sector organizations, universities, think tanks and other entities. The aim was to expand the focus to as wide a range of documentary sources as possible. The summary of results included 60 documents from this search.

- **Semi-structured interviews with AI and algorithmic evaluation experts.**

We conducted 15 interviews with experts working in international and European organizations, private companies and consultancies, universities and third sector organizations – ten in English and five in Spanish. The perceptions of the specialists and experts interviewed complemented the field research and helped identify aspects and bring clarity to issues less addressed in the academic literature.

With these sources we conducted an in-depth examination of a topic of major importance to contemporary societies. It will be of particular interest to political representatives, staff of public and third sector organizations, activists and specialists in the field, as well as the general public. The debate on algorithm evaluations invites us to reflect on the impact of AI on people and organizations, and on the future relationship between people and machines.

1. Why evaluate algorithms in the current context of artificial intelligence?



AI algorithms and systems are sparking mixed and often contradictory reactions across growing numbers of industries. Pessimistic and optimistic positions on AI and the future of humanity coexist, as seen in the public and political discourse, the media and studies of individual perceptions. Some experts believe AI systems will aid human progress effectively, while others argue they may cause harmful societal changes (Stanford University Human-Centered Artificial Intelligence 2023).

As past technological advancements have shown, we cannot precisely predict how new systems will affect the future. But we can get a head start by analysing how algorithm-based AI systems work, what real impacts they are having now and what risks they entail for the future. This section assesses the need to evaluate algorithms at a time when there is certain consensus regarding the beginning of the so-called Fourth Industrial Revolution.

The field of algorithm evaluations offers certain conceptual and practical answers to these concerns. In recent years, debate has been mounting in academia, civil organizations and governments regarding the current relevance of algorithmic evaluations, and the most appropriate methods and tools to implement them. So why is it important to conduct algorithmic evaluations in public and private organizations? These are some of the arguments:

- **The use of AI and algorithms is having a tangible impact.** While the future with AI is uncertain, there is evidence of its potential and current negative impacts on particular populations. AI data and models can contain gender, racial and other biases that result in discrimination against certain groups (Buolamwini and Gebru 2018; Morondo and Eguiluz 2022). There are also cases of algorithms that have been used to automate processes inappropriately with harmful consequences to vulnerable people, such as exclusion from certain public services and social care (Eubanks 2019).

AI systems have also been noted to have a high environmental impact due to their hardware and server requirements, as well as their enormous consumption of resources (raw materials, electricity, etc.) and cloud storage time (Strubell et al. 2019). Hence, there are already compelling reasons to evaluate the possible negative effects of their design and implementation.

- **Evaluations are in part new – but not entirely.**

Algorithm evaluations are based on the long-standing tradition of audit and evaluation processes in other fields. Financial audits, for instance, are a case in point, but there are also models for carrying out social and ethical impact assessments, as well as impact assessments on privacy, data protection and human rights (Mantelero 2018). These frameworks for analysis provide a valuable roadmap for examining the functioning and impact of algorithmic systems on society, though they need some adaptations.

Given the complexity of AI and its growing implementation in a wide range of scenarios, it is essential to be armed with strategies to evaluate the particularities of these systems. The limitations in their processes need attention limitations such as the difficulty of identifying certain harms and the risk of an evaluation being reduced to a checklist of indicators that is given no further reflection (Mökander et al. 2022). Nevertheless, with the right perspective, and drawing on lessons learned in other sectors, algorithmic evaluations have enormous potential to identify and mitigate the negative impacts of algorithmic systems.

- **Evaluating the use of algorithms is a growing legal and ethical obligation.**

Currently there is little regulation of algorithmic evaluations with a few exceptions, such as New York City Local Law 144 or the Canadian Government's Automated Decision Directive. The models of the European Union, North America and China all have significant differences, which will likely lead to differentiated developments. The EU is leaning towards a more regulatory approach, as shown by the AI Law soon to come into force.¹ Meanwhile, in the English-speaking arena, there is a tendency to avoid excess regulation, so algorithm evaluations are likely to take the form of ex post certifications or codes of conduct.² In China, the decisive role of the state and a social culture anchored in Confucianism could see the process take place under the watchful eye of the government,³ with other stakeholders taking a secondary role.

¹ At the end of 2023, the Presidency of the European Council and the European Parliament reached a provisional agreement on the future Artificial Intelligence Act. The draft regulation aims to ensure that AI systems used in the EU are safe and respect fundamental rights and EU values (<https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>).

² However, in 2022 the proposed Algorithmic Accountability Act was introduced in the US Senate. This would require large companies to develop impact assessments of their algorithms (<https://www.congress.gov/bill/117th-congress/senate-bill/3572>).

³ The Chinese government launched a mandatory register of recommended algorithms, in which companies, in addition to including general information on the system, are required to upload a document with a security self-assessment. The criteria for understanding security risks lie solely with the government and the evaluation processes are not open to the public (Sheehan and Du 2022).

For Ricardo Baeza-Yates, research director of the Institute for Experiential AI at Northeastern University in the United States, looking at democratic countries there is a clear distinction between those with greater confidence in their institutions and those where institutions are not perceived to function properly. In the first case, a model favouring accountability (i.e. conducting evaluations to establish responsibilities once the algorithms have been implemented) would be apt; in the second, a focus on transparency would hold more weight (that is, making certain information public throughout the AI life cycle).⁴

Despite this difference in models, evaluations are considered indispensable mechanisms to ensure algorithmic accountability (Basu et al. 2021). Whatever form they take (whether under strict regulations or as voluntary mechanisms), governments and/or civil society are expected to increasingly require algorithmic evaluations.

- **Algorithmic evaluation is a socio-technical issue.**

The increasing prevalence of algorithmic systems across different areas of daily life requires a shift in focus from solely technological aspects to the broader public and social concerns when assessing their implications. This is particularly significant when examining different ethical approaches to algorithm use, especially in terms of equality and non-discrimination. For instance, the technology sector has often understood that efforts should focus on identifying and mitigating biases in data and models, a predominantly technical approach that does not necessarily address the underlying issue of structural discrimination (Morondo and Eguiluz 2022, p. 27).

That is, the meanings of bias and discrimination in the technological sphere do not necessarily accord with the concepts as they are understood in other disciplines, and which require special focus. Considering the advances of AI and its impact on the daily lives of increasing numbers of people, it is crucial to move beyond a purely technical viewpoint to consider the interplay of technology with human and social dimensions.⁵ This socio-technical perspective urges us to broaden algorithmic evaluations and apply a holistic approach to address a multifaceted reality with a focus on people.

⁴ Transparency and accountability are interrelated, but in some cases one is given more prominence than the other, though transparency requirements are an essential aspect and very present in all countries.

⁵ For Javier de la Cueva, patron of the Civio Foundation and specialist in law, information technology and communication, it is impossible to separate the technical aspects from the political, social and cultural dimensions. That is, certain social, political and cultural constructs manifest in the way the technologies are designed and implemented.

2. Conceptual framework: definitions and dimensions of algorithm evaluations



Having recognized the importance of evaluating algorithms, it is necessary to clarify certain concepts. For instance, what exactly is an algorithmic evaluation? Defining it is challenging as there are many interpretations across academic studies, and reports of civil society organization and government agencies.

Audits or impact evaluations are generally defined as mechanisms used to identify problematic behaviours in algorithmic systems (Bandy 2021), but with emphasis on different objectives, such as detection of bias and discrimination in algorithmic decisions (Minkinen et al. 2022; Sandvig et al. 2014);⁶ assessment of potential harm (Baykurt 2022); analysis of risk levels in terms of human rights, ethics and privacy (Yam and Skorburg 2021); or the study of the impact on the rights and interests of certain groups (Brown et al. 2021). Some definitions not only emphasize identifying the problems, but stress the need to point out possible mitigation solutions and strategies.⁷

⁶ Recognizing the importance of addressing discrimination and algorithmic biases requires acknowledging the diverse approaches available. As highlighted by Morondo and Eguiluz (2022), biases in an algorithmic system's data and models can impact its functionality, leading to the implementation of predominantly technological mitigation measures. However, this perspective may overlook the structural discrimination experienced by certain populations. There is hence a need to pivot towards a comprehensive and multidisciplinary approach that considers all facets of algorithmic discrimination and bias.

⁷ Adriano Soares Koshiyama, co-founder of the company Holistic AI, believes that specialists in algorithmic evaluations must play their role similarly to healthcare staff: that is, go beyond diagnosis, and offer solutions.


Terminology matters

While understanding of the terminology surrounding algorithm evaluations varies (Ada Lovelace Institute 2020), this document uses the term *evaluation* to refer to the overall process of analysis of algorithmic systems and identification of problems. We distinguish between the umbrella idea of *evaluation* and the concept of *impact assessment*, which is considered a specific methodology for evaluating algorithms and which will be explained later (Ibid.).



A clear distinction exists between the terms *evaluation* (evaluation in its broad sense), *impact assessment* (impact assessment as a specific methodology) and *audit* (focused on meeting technical or non-technical requirements). We will address further related concepts, such as *algorithmic accountability*, later.

Beyond this broad conceptualization, certain distinctions are needed. Not all algorithmic evaluations are the same or cover the same elements. Based on our study of the academic literature, documentary review of official government sources and agencies dedicated to audits and evaluations, as well as our interviews with experts, we propose an algorithmic evaluation with ten dimensions: focus, locus, stakeholders, role of external actors, timing, regulatory focus, topic, scope, level of access and methodology. Table 1 summarizes this evaluation type, presenting a set of questions that aids in gaining a clearer understanding of each dimension.

Table 1.
Dimensions of algorithm evaluations

▼ Dimension	Categories	Reference questions
 Focus	<ul style="list-style-type: none"> • Technical approach • Holistic approach 	<p>Is there special access to training data, the model, outputs or other technical aspects of the system?</p> <p>Is there access to information that facilitates analysis of the relationship between technological elements and the social, organizational, cultural and contextual aspects?</p>

 Dimension	Categories	Reference questions
 Locus	<ul style="list-style-type: none"> • Internal • External 	Is the organization implementing the algorithmic system involved in the evaluation process? Is there express authorization and access to internal information regarding the development of the algorithm?
 Promoting actors	<ul style="list-style-type: none"> • Primary (first-party) • Secondary (second-party) • Tertiary (third-party) 	Who leads the algorithmic evaluation process?
 Role of external stakeholders	<ul style="list-style-type: none"> • Participatory • Non-participatory 	Which stakeholders are involved in the algorithmic evaluation process? Are the users and groups potentially affected by the algorithmic system involved? If the affected users and people are involved, what form does their participation take?
 Timing	<ul style="list-style-type: none"> • Ex ante • Ex post 	Does the algorithmic evaluation process take place before or after the implementation of the system?
 Regulatory focus	<ul style="list-style-type: none"> • Legal obligations • Compliance with regulatory frameworks • Good practices • Certifications 	What is the purpose of developing the algorithm evaluation? Is it intended to comply with specific regulations, promote good practices on a voluntary basis, or obtain a quality certification, etc.?
 Topic	<ul style="list-style-type: none"> • Use of data • Ethics and human rights • Governance 	What is the main topic of the evaluation process? Is the priority the analysis of data use, ethical and human rights aspects, or system governance?
 Scope	<ul style="list-style-type: none"> • Specific aspect of an algorithm • Complete system 	What is the purpose of the analysis? Is it a specific aspect of the system (data, model, etc.) or is it intended to gain understanding of its entirety?

▼ Dimension	Categories	Reference questions
 Level of access	<ul style="list-style-type: none"> • White-box • Black-box 	What level of access to the algorithmic system does the evaluation team have? Is there unrestricted access to all the information or are there limitations to obtaining internal data?
 Methodology	<ul style="list-style-type: none"> • Auditing • Impact assessments 	What specific methodology is followed in the evaluation process? Is it intended to analyse the algorithm based on a series of specific criteria or is it intended to understand its potential risks or impacts?

Data source: own creation based on Ada Lovelace Institute 2020; Costanza-Chock et al. 2022; Meßmer and Degeling 2023; Metcalf et al. 2021; Kelly-Lyth and Thomas 2023; Koshiyama et al. 2021 and interviews with subject experts



Focus

The focus can adopt either a technical or a holistic approach. In a decidedly technical approach, the aim is to understand the functioning of the algorithm and/or codes transforming inputs into outputs. The results of and decisions made by the algorithm are evaluated according to specific, technical criteria to identify any biases in the algorithm's data and models.

One notable case is the research of Buolamwini and Gebu (2018), which audited classification systems and found gender and racial biases. The findings were published in the widely-recognized article *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* (2018).

The holistic approach involves identifying issues from a broader standpoint, going beyond the functionality of algorithms to consider the context, structures, stakeholders and other factors that interplay with the deployment of algorithms and shape their outcomes. While there is a clear distinction, the optimal strategy is to integrate both approaches, ensuring the evaluation process is as comprehensive as possible.

One interesting example halfway between the two approaches is the work of Papakyriakopoulos and Mboya (2023). They developed a socio-computational method for analysing biases in the Google image search engine by combining technical analysis of the system with the use of critical theories of power.



Locus

Algorithmic evaluations can be either **internal** or **external**. In the former, the process takes place within a specific organization and is usually carried out by internal staff. External evaluations do not require the participation or authorization of the organization under analysis. This process takes place in external independent settings (universities, media, etc.).

One of the best-known examples of an external evaluation is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) system developed by the independent media organization ProPublica in 2016. Their data analysis revealed racial bias in predictions of recidivism in criminal behaviour.



Promoting actors

Closely related to the previous dimension, this one identifies three types of evaluations according to the stakeholders driving the process (Costanza-Chock et al. 2022; Meßmer and Degeling 2023; Metcalf et al. 2021). The audit or impact assessment may be a **first-party** initiative, i.e. by staff of the organization itself (the internal locus). Generally, the findings of these evaluations are not made public (Costanza-Chock et al. 2022). One example is the process Amazon conducted internally to evaluate the algorithm used in staff selection, which detected gender bias (Dastin 2018).

Organizations may hire **second-party stakeholders**, i.e. external organizations (usually consultancies, foundations, etc.) to conduct second-party evaluation processes. The professionals who carry out these evaluations are not fully independent, as they comply with indications from their hiring organizations in one way or another. Examples of such actors are O’Neil Risk Consulting & Algorithmic Auditing, the company founded by scientist Cathy O’Neil to evaluate algorithms, and HireVue, which is used for staff recruitment.

Finally, **third-party evaluations** are conducted with full independence from the organization evaluated. Independent supervisory organizations, researchers from academic institutions or journalists evaluate the algorithmic systems and publish their analyses to raise awareness across society. Efforts have also been made to empower members of the public without technical expertise to assess the algorithms that affect them. The IndieLabel tool, for instance, was designed so that users could train a model and easily identify toxic comments on content platforms (Lam et al. 2022).



Role of external stakeholders

Algorithmic evaluations can be understood according to the different roles of the communities affected and the general public. On the one hand, **non-participatory** evaluations are conducted by specialists, external or internal, without considering other actors external to the evaluation process who could contribute their viewpoint of the algorithmic system.

Meanwhile, an increasing number of experts stress the importance of adopting a **participatory** approach. They argue that this method improves diversity in evaluations, increasing confidence in the process and contributing to genuine accountability (Groves 2022). The focus is on consulting and involving those affected by algorithmic decisions, such as civil society organizations, entities with specific interests and users. This inclusive approach helps to detect any potential undesired effects or unacceptable impacts, while broadening the evaluative perspective, especially, though not exclusively, among those potentially most affected by the algorithmic decisions.



Timing

The timing of algorithmic evaluations is a key dimension. Several studies highlight the importance of ongoing evaluations throughout the life cycle of an AI system (Mökander et al. 2022; Novelli et al. 2023; Sandu et al. 2022). Evaluations can take place *ex ante* or *ex post*. **Ex ante** algorithmic evaluations are conducted before an algorithm is implemented, addressing key questions regarding the assumptions behind its design and potential risks (Ada Lovelace Institute 2020; Sloane 2021). Following implementation, **ex post** evaluations are needed to comprehend the real-world impacts of its use (Ada Lovelace Institute 2020; Eticas Consulting n.d.), including any that were unforeseen. Real-time evaluations, which take place during the implementation of the algorithmic system, also contribute to this understanding.



Regulatory focus

Algorithmic evaluations can also be classified according to their degree of regulatory constraint or focus. The process is aligned with regulatory inspections, explained later in the methodology section. Building on the basis established by Kelly-Lyth and Thomas (2023), who interpret Burr and Leslie (2022), we propose three categories depending on the extent and type of regulatory compliance of the organizations involved.

The first category, **legal obligations**, refers to the algorithmic assessment processes developed to comply with the regulations of a given geographic context. For instance, an evaluation may be developed to ensure compliance with Article 22 of the General Data Protection Regulation, which states that any data subject has the right “not to be subject to a decision based solely on automated processing, including profiling”.



Second are the evaluation and audit processes designed to comply with a **broader and less binding regulatory framework**, such as documents or agreements with general principles or recommendations regarding AI, such as the European Charter on the Ethical Use of Artificial Intelligence in Legal Systems, or the UNESCO Recommendation on the Ethics of Artificial Intelligence, among many others. In other cases, the intention is to voluntarily analyse whether the use of algorithms aligns with various **good practices**, such as contributing to gender or ethnic equality or compliance with human rights. Finally, the option of **certifications** notably offers a quality seal to organizations that meet certain ethical standards in the design and implementation of the algorithms (De Manuel et al. 2023).



Topic

Under certain circumstances algorithms may be evaluated in the context of a highly specific topic. While numerous topics exist, three key areas have been highlighted in various papers: data use, ethics and human rights, and governance. Thus, it may be of interest to focus on the **data** used in algorithmic systems, specifically in terms of privacy, transparency and protection of personal data.⁸ A broader evaluation of algorithmic system **governance**, without focusing on any specific issue, could also be viable. Another option is to develop a process to assess **applied ethics and compliance with human rights**, incorporating a set of well-defined indicators.

In the latter case, it is important to adapt to each context. Sherry Wasilow and Joelle B. Thorpe (2019), for example, propose an ethical assessment framework for AI and robotics systems in the Canadian military, including compliance with country-specific codes of ethics and regulations, as well as considerations of health, safety, equality, trust, security, human dignity and others. Another notable example is the Human Rights, Ethical and Social Impact Assessment-HRESIA (Mantelero 2018), a tool that combines a self-assessment questionnaire and the perspective of a committee of specialists (when necessary) to analyse both ethical and human rights aspects of AI systems.



Scope

The scope of an algorithm evaluation can also vary according to its objectives, interests and resources. Some evaluations focus on a **specific aspect** of an algorithm, such as its training data, its underlying model or its expected results, among other things (Garde Roca 2023). Evaluations can also cover the **entire life cycle** of a system and other broader facets of the context in which it is deployed.

⁸ In 2021 the Spanish Data Protection Agency published a document entitled: [Requisitos para Auditorías de Tratamientos que incluyan IA](#) (Requirements for Processing Audits that Include AI), which identifies a series of data protection controls in processing using AI components.



Level of access

Depending on the stakeholders involved, locus, and other factors, there may be greater or lesser access to the data needed to conduct an algorithm evaluation. A paper by Koshiyama et al. (2021) explains that there are **seven levels of access**: at one end are **white-box** processes (number 7 on the scale), in which it is possible to obtain all system details. At the other end are **black-box** evaluations (number 1 on the scale), in which “only indirect system observations can be made” (Koshiyama et al. 2021, p. 4). There is also an **intermediate zone** as there is a progressive decline in the level of access between numbers 7 and 1.

Notably this classification refers specifically to the algorithmic evaluation process and not to the use of the algorithmic system. In other words, the individual or team responsible for the evaluation may have full access to an algorithmic system, even if it is considered a black-box system because its data and functioning are not open to the public, in which case it would be a white-box evaluation. In this context, although the algorithm is not publicly accessible, whoever conducts the evaluation has privileged access to the system allowing for a comprehensive evaluation.



Methodology

Algorithmic evaluations follow different methods. As noted, the term *evaluation* generally refers to analysis of the use of algorithms. But when focusing on specific methodologies, certain conceptual distinctions are essential. The terms *audit* and *algorithmic impact assessment* are often used synonymously yet differences underlie their methodologies.

One of the most important reports on the subject, by the Ada Lovelace Institute (2020), notes that **audits** focus on analysis of the functioning of an algorithm in relation to specific criteria, such as specific assumptions of bias (bias audits) or standards set out in regulations (regulatory inspections). In this case, the process is conducted after the implementation of the algorithmic system once its effects in specific contexts have been identified.

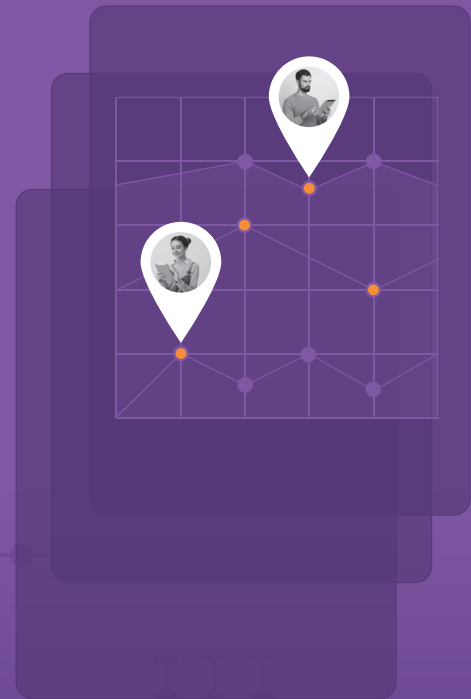
Algorithmic impact in contrast, take a broader approach. They can measure the risks involved in a system among certain groups of people before or during implementation (*algorithmic risk assessment*) or measure their impacts after implementation (*algorithmic impact evaluation*).

Table 2.
Specific methodologies for assessing algorithmic systems

Audits		Impact	
Bias audit	Regulatory inspection	Risk assessment	Impact evaluation
Analysis of specific hypotheses of bias that may exist in the data or models	Analysis of compliance with regulatory standards	Measures the risks to certain groups (conducted before or in the early stages of implementation)	Measures the impact of the algorithm after its implementation

Data source: Ada Lovelace Institute 2020

3. Algorithm evaluation methods























In the wide range of approaches to algorithmic evaluations, there has been progress in the design and application of a variety of methods that address the particularities of the process. Each approach offers advantages and disadvantages, and the suitability of each depends on the proposed objectives and available resources. This section addresses the existing methods and connects them with the dimensions detailed in the previous section. This completes the map of the state of the issue, which is intended to offer new knowledge and guide next steps to improving algorithmic analysis processes.

The academic literature and official documents and reports of various organizations identify various methods of algorithmic evaluations. The most common methods are code audits, scraping, sock puppet, carrier puppet, collaborative audits, statistical analyses, checklists, user surveys, workshops or focus groups, and case studies or development histories.

The table below is intended to bring clarity to the debate and enable a more detailed understanding of these methods (see Table 2). As it shows, most methods are technical and non-participatory, which opens the door to expanding perspectives towards other more qualitative and holistic strategies in future (Bandy 2021; Costanza-Chock et al. 2022). It shows the need to continue to expand the frontiers of knowledge in this multifaceted field with an approach that is both social and technical.

Table 3.
Methods and techniques to evaluate algorithms and their relationship with their identified dimensions

▼ Dimension	Code audits	Scraping	Sock puppet	Carrier puppet	Collaborative audits
 Focus	Technical	Technical	Technical	Technical	Technical
 Locus	Internal (more likely) or external	Internal or external (more likely)	Internal or external (more likely)	Internal or external	Internal or external
 Promoting actors	First-party (more likely), second-party (more likely), or third-party	Second or third-party	First-party, second-party or third-party (more likely)	First-party, second-party or third-party (more likely)	First-party, second-party or third-party
 Role of external stakeholders	Non-participatory	Non-participatory	Non-participatory	Non-participatory	Between participatory and non-participatory
 Timing	Ex ante or ex post	Ex post	Ex ante or ex post	Ex ante or ex post	Ex ante or ex post
 Regulatory focus	Legal obligations, compliance with regulatory frameworks and good practices	Good practices	Legal obligations, compliance with regulatory frameworks and good practices	Legal obligations, compliance with regulatory frameworks and good practices	Legal obligations, compliance with regulatory frameworks and good practices
 Topic	Ethics, human rights or governance	Ethics, human rights or governance	Ethics, human rights, or governance	Ethics, human rights or governance	Ethics, human rights or governance
 Scope	Specific aspect	Specific aspect	Specific aspect	Specific aspect	Specific aspect
 Level of access	White-box	Black-box	Intermediate or black-box	Intermediate or black-box	Intermediate or black-box
 Methodology	Audit	Audit	Audit	Audit	Audit

▼ Dimension	Statistical analyses	Checklists	User surveys	Workshops or focus groups	Case studies or development histories
 Focus	Technical	Technical and/or holistic	Holistic	Holistic	Holistic
 Locus	Internal or external	Internal or external	External	Internal or external	Internal or external
 Deployers	First-party, second-party or third-party	First-party (more likely), second-party (more likely) or third-party	First-party, second-party or third-party (more likely)	First-party, second-party or third-party	First-party, second-party or third-party (more likely)
 Role of external stakeholders	Non-participatory	Non-participatory	Participatory	Participatory	Participatory or non-participatory
 Timing	Ex ante or ex post	Ex ante or ex post	Ex post	Ex ante or ex post	Ex post
 Regulatory focus	Legal obligations, compliance with regulatory frameworks and good practices	Legal obligations, compliance with regulatory frameworks and good practices, certifications	Good practices	Good practices	Good practices
 Topic	Data use, ethics, human rights or governance	Data use, ethics, human rights or governance	Ethics, human rights or governance	Ethics, human rights or governance	Ethics, human rights or governance
 Scope	Specific aspect	Entire system	Entire system	Entire system	Entire system
 Level of access	Intermediate	White-box, intermediate or black-box	White-box, intermediate or back-box (more likely)	White-box, intermediate or back-box	White-box, intermediate or back-box
 Methodology	Audit	Audit or impact evaluation	Impact evaluation	Impact evaluation	Impact evaluation

Data source: own data based on Koshiyama et al. 2021; Hamilton 2021; Raji and Buolamwini 2019; Pappu et al. 2021; Sandvig et al. 2014; Oswald et al. 2018; Wasilow and Thorpe 2019; and interviews with experts

Code audits

The purpose of code audits is to move towards greater algorithmic transparency (Sandvig et al. 2014). Code audits are primarily technical audits in which source code is analysed for potential issues such as discrimination against certain populations or privacy concerns. It is usually run internally at the initiative of the organization concerned. This is because, as noted in the literature (Koshiyama et al. 2021; Sandvig et al. 2014), it requires full access to information that may be sensitive (white-box), which organizations are unlikely to disclose publicly.

If an external organization is contracted for a second-party audit, information may be shared with them to conduct a comprehensive audit. This technical audit is highly specific, follows predefined parameters and typically does not involve the participation of external stakeholders due to its focused nature. It can be conducted before or after implementation of the algorithm, either to fulfil legal obligations (if specified in the regulations) or voluntarily to comply with standards and good practices.

Use cases - Code audits

- One of the best-known tools for auditing machine learning models is AI Fairness 360 (<https://ai-fairness-360.org/>). An open-source software initially developed by IBM and currently under the initiative of The Linux Foundation, it can detect biases in various aspects of the AI cycle, including the algorithm itself. Another known tool used to identify algorithmic biases is Aequitas, developed by University of Chicago researchers (Saleiro et al. 2019).
- In 2020 Pymetrics hired a team from Northeastern University in the United States (Wilson et al. 2020) to audit its algorithmic job application assessment tool. The researchers had access to the source code and additional documentation and found no discriminatory results, according to very specific parameters. This case is not exempt from criticism due to the limitations of the baseline definitions used to measure discrimination (Schellmann 2021).

Scraping

This type of evaluation seeks to interact intensively with the algorithm to evaluate its performance and results. The researcher can make manual requests (to evaluate a search engine algorithm for instance), but different to how conventional users do so (Pappu et al. 2021; Sandvig et al. 2014). They can also use APIs (i.e. application programming interfaces, which facilitate communication and interaction between different software systems and components to obtain the desired information). It is a technical evaluation, but, unlike in code audits, the source code is not fully accessible so it can be conducted externally by contracted organizations or independent consultants.

Scraping is generally used after the algorithm is implemented and without the participation of affected communities or people. These processes are unlikely to be legally required, so they may rather be understood to promote good practices in terms of ethics, human rights or other issues of interest.

Use cases - Scraping

- Orestis Papakyriakopoulos and Arwa Michelle Mboya (2023) developed an ingenious socio-computational framework using scraping methods to evaluate gender and racial biases and stereotypes in the Google image search engine. They trained a program with tagged images, which automatically sent queries to the Google search engine and extracted the answers from the algorithm. They then used computational methods in combination with qualitative analysis to interpret the results.
- In a 2018 investigation, Kulshrestha et al. (2019) analysed political biases in Twitter and Google search results in relation to the 2016 United States presidential primaries. To achieve this, they interacted directly with the platforms and obtained data that was publicly and openly available.

Sock puppet

Sock puppet programs act as system users and evaluate the decisions an algorithm makes based on their profile. Unlike scraping, sock puppet evaluations retrieve more detailed information on the specific variables studied (Pappu et al. 2021), though the ethics of the method are in dispute (Sandvig et al. 2014).

An organization may conduct this type of evaluation internally to detect possible problems with the algorithm before or after its implementation, and to comply with specific laws or encourage good practices. However, it is more likely to be used at the request of a contracted organization or at the behest of independent researchers. It is an option when there is no detailed information on the source code (so it can be considered black-box or an intermediate stage of the access spectrum), and the participation of affected communities in the process is not necessary.

Use cases - Sock puppet

- Researchers Eriksson and Johansson (2017) created 288 Spotify accounts (bots), half registered as men and the other half as women. Their intention was to check for gender bias in the platform's music recommendations.
- Recently, a group of researchers (Srba et al. 2023) used the sock puppet method to study the risks of falling into a misinformation bubble filter on YouTube. They programmed bots to put themselves in the place of platform users, and analysed the searches, the homepage results and the video recommendations.

Carrier puppet

This method is similar to the sock puppet method except that the program acts as the developer rather than the end user. That is, tests are run to detect possible problems at an intermediate stage of system development, not with the end product (Raji and Buolamwini 2019). Its characteristics and possibilities are very similar to the previous case.

An organization may wish to carry out this type of evaluation before launching a product, but it is much more likely to be done at the request of external stakeholders to promote good practices and raise awareness of ethics, human rights and other issues. It does not involve the participation of the communities directly affected, nor is it necessary to obtain full information on the code, though some degree of access is needed to conduct the process properly.

Use cases - Carrier puppet

- The best-known carrier puppet case is the paper by Buolamwini and Gebru (2018) called Gender Shades, in which this method was used to detect gender and racial biases in facial recognition systems. The researchers found that these systems are less successful in identifying black women so they can have a harmful effect on this community.

Collaborative audit

Collaborative audits are similar to sock puppet audits with the difference that users are hired to test the system (Sandvig et al. 2014). This method can be used both internally and externally, either at the initiative of the organization itself (with internal or external staff) or in the interests of independent researchers. When the approach is internal, it is done before the algorithm is deployed to verify any problematic behaviour, or after implementation, especially if carried out by an external evaluation organization. If it is conducted at the initiative of the organization itself, it can be a good strategy to use to comply with legal obligations or regulatory frameworks, while promoting good practices.

Some clarifications are required regarding participation. This type is classified as intermediate – between participatory and non-participatory – because, while it includes algorithm users, the designs are usually experimental with each group being required to follow specific instructions. That is, user experiences are considered, but the people affected or potentially affected by the algorithm are not necessarily included, unless the design and approach of the audit so establishes.

Use cases - Collaborative audit

- In a study by Spyridou et al. (2022), 18 people participated, divided into two groups. Each participant installed a plug-in in their search engine and interacted with the MyNews portal according to specific instructions. The information collected was used to analyse the behaviour of the news recommendation algorithms.
- Independent journalism portal The Markup developed the Citizen Browser project to audit social media algorithms, specifically Facebook. A total of 1,000 United States residents were paid to install on their personal computers an application that collected information on their use of the social network. Research using the data collected has been published on abortion, cryptocurrency-related scams, extreme right content, and other topics (<https://themarkup.org/series/citizen-browser>).

Statistical analysis

Statistical analysis of system data and results is another method frequently used in algorithm evaluations. This process can be conducted both internally and externally, at the initiative of the organizations themselves, by external consultants or by independent evaluators. It requires access to certain data so it can be framed at an intermediate level (between white-box and black-box). For instance, if an organization has relevant data, it is feasible for independent investigators to use this data to perform statistical analysis on certain variables of interest (Hamilton 2021). The organizations may use the data to carry out this type of study internally, before or after putting an algorithm into use.

This process facilitates making inferences on a specific aspect of a system (bias, discrimination, privacy, etc.), but it is more limited than code or model analysis. As in other methods, the involvement of outside communities is not needed, and it can be used as a complement to other methods to meet regulatory obligations or to voluntarily promote the ethical deployment of AI in the organization.

Use cases - Statistical analysis

- The Model Risk Audit proposed by Munz et al. (2023) includes statistical analysis to evaluate AI models in four categories: 1) robustness; 2) security and privacy; 3) explicability and bias; and 4) performance and methodological integrity. It is particularly important in the case of explicability and bias.
- Professor Melissa Hamilton of the University of Surrey in the UK published the study, Public Safety Assessment (Hamilton 2021), on an algorithmic tool used in the United States to make predictions in the context of preliminary research (before going to trial). One of the aspects the system is intended to predict is the risk that a person under investigation will not show up for court appointments before a final decision has been made on their case. To evaluate the algorithm, the researcher statistically analysed the predictions of the tool and the actual case data in three states in the United States.

Checklists

Checklists are used to collect relevant information on algorithm use in relation to a series of predefined indicators. They may include open or closed questions (usually with yes/no answers), or tables built to collect specific information – there are many options. These instruments are used internally to run checks before launching an algorithm, or an external consultant can run the process.

Independent researchers or supervisory bodies can use checklists to gather the information they need to evaluate an algorithm, provided the organization collaborates. Some cases may require interviews with developers, staff of the organization, etc., or the use of technical tools to analyse the details of the algorithm. Depending on how it is designed, it can facilitate study of the entire system and its relationship with the context in which it is deployed, with a focus on a specific topic (data use, human rights or environment, for instance) or with more general questions.

Use cases - Checklists

There are several tools for developing algorithm monitoring checklists. Some are listed below:

- Algo-care guidance framework (Oswald et al. 2018)
- Human Rights, Ethical and Social Impact Assessment-HRESIA (Mantelero 2018)
- Ethics Assessment Framework (Wasilow and Thorpe 2019)
- After-Action Review for AI (Dodge et al. 2021)
- Government of Canada Algorithmic Impact Assessment Tool (<https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html>)
- Algorithmic Impact Assessment Tool of the Chief Information Officers Council

User surveys

Surveys of real users, also called non-invasive audits (Sandvig et al. 2014), can be useful to learn the impact of an algorithm that has already been deployed. The information collected on user perceptions is used to make inferences regarding the system's functioning, though no causal relationships can be established between the variables studied (Ibid.). Surveys can be used as a complement to other more technical procedures to obtain information regarding the actual system operation.

When other methods are not possible, surveys are useful to obtain data and gain a holistic understanding of an algorithm's real-life impact. This means that while an organization can promote such processes to discover an algorithm's impact, any independent evaluator can also conduct this type of study to raise awareness of the subject. It also gives the chance to raise queries on various topics, such as ethics, human rights and the environment, among others.

Use cases - User surveys

- A team of researchers from Stanford and Pennsylvania Universities and the Georgia Institute of Technology (Lam et al. 2023) designed the Intervenr platform to conduct socio-technical evaluations on Internet search engines. As part of their research, they complemented user behaviour observation (which would amount to a collaborative audit) with surveys of their experience.

Workshops or focus groups

The academic world is beginning to advocate for the incorporation of real users into algorithmic evaluation processes using qualitative research methods (DeVos et al. 2022; Groves 2022). In addition to interviews, one of the options proposed is workshops, in which participants can express their ideas and experiences of the impact of algorithms.

Workshops offer a range of possibilities. An organization might organize a meeting with people who may be affected by a system, allowing them to freely test it and share their insights before its implementation. Or an independent consultant or researcher might run a workshop to evaluate an algorithm already in use. Workshops and focus groups can bring a more holistic approach to evaluations, enabling exploration of a variety of topics from different perspectives.

Use cases - Workshops or focus groups

- Researcher DeVos et al. (2022) used a combination of think-aloud interviews (gathering opinions while using an algorithmic system), diary studies, and workshops to gain the user perspective in algorithmic impact evaluations.
- A research team composed of the consultancy Eticas, the Pompeu Fabra University of Barcelona and ALPHA Telefonica evaluated the REM!X app, developed by Telefonica Innovation Alpha, to offer recommendations for well-being (Galdon Clavell et al. 2020). Their evaluation followed four strategies: analysis of algorithm recommendations, document review, digital ethnography (a type of research that analyses the social relationships that occur in the online environment) and study of feedback messages from users, with five focus groups in which evaluators participated, as did the app developers and engineers.

Case studies and development histories

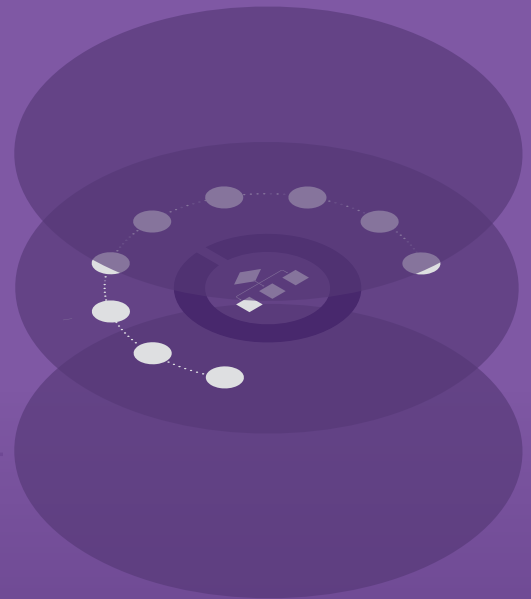
Case studies and development histories can be used to address the complexity of algorithmic systems (Bandy 2021) with an approach similar to digital ethnography. The first option involves in-depth analysis of a specific case, preferably using direct observation, interviews and other qualitative methods to capture the organizational and socio-technical dimensions of the algorithms.

The second option is about reconstructing the history of the development of an algorithm to understand the source of its problems and find potential solutions. As in the previous case, these evaluations can be carried out at the initiative of the organizations or by external actors but are more likely to be done by independent consultants or researchers. Depending on the approach, they may involve the communities affected or focus solely on the staff of the organizations that design and implement the algorithms. However, to obtain all the information needed, conducting an evaluation after the system is implemented is preferable.

Use cases - Case studies and development histories

- DeVito (2017) analysed Facebook press releases, patents and official documents to reconstruct the development history and identify the main values in the algorithm that could explain the content selection in Facebook's news feed.
- The Eticas consultancy conducted an external audit of the Viogén system used in Spain to predict the risk of a woman becoming a victim of gender violence a second time. The audit team did not have access to the original data used but used secondary data for statistical analysis, interviews with 31 women and a questionnaire with seven lawyers. While the perspective of the public personnel using the system was not available, it was possible to reconstruct its main characteristics and limitations through the testimonies of women who had experience with this tool during their cases.

4. The ecosystem of algorithmic evaluations and governance levels of algorithmic accountability



Algorithmic evaluation processes are not conducted in isolation. They form part of a broader ecosystem with various levels of governance, with the participation of different spheres (public, private and social), sectors or areas of activity (health, education, transport, banking, energy, energy, security, etc.) and stakeholders (both directly and indirectly related to the process). An algorithmic evaluation ecosystem is taking shape with a number of interacting components that relate to the broader AI and accountability landscape, both at the national level of each country and internationally (Percy et al. 2021; Stahl et al. 2023).

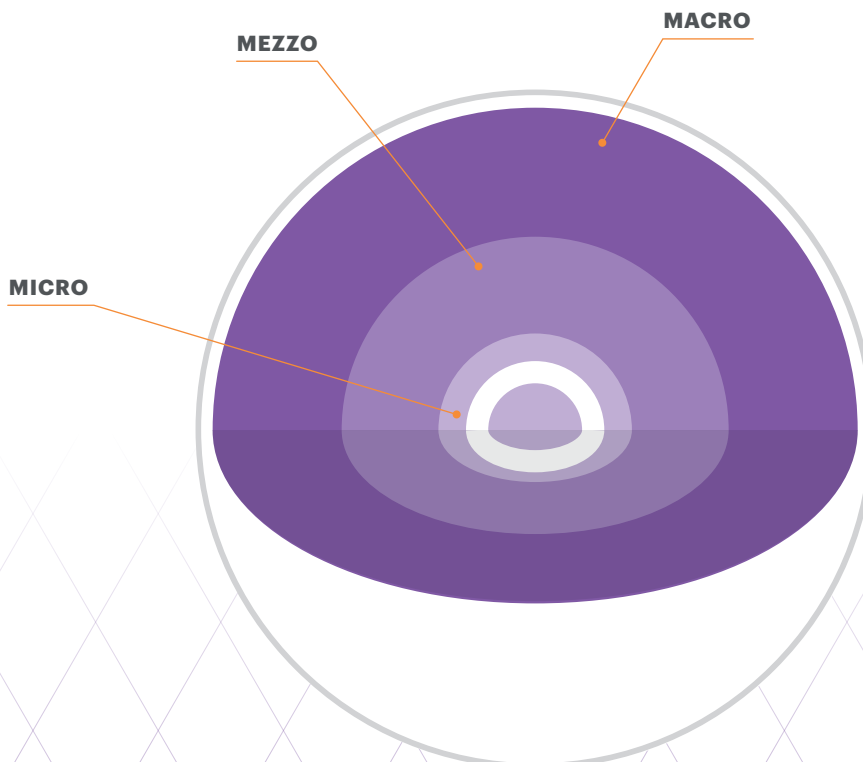
While there is still a long road ahead, Costanza-Chok et al. (2022) point out that this ecosystem is undergoing rapid growth and as such it deserves the attention of regulators, developers, companies, consultants, public administrations and academia, but also of system users and civil society as a whole, to promote specific standards (Costanza-Chok 2022) that contribute to the development of commonly accepted practices, considering the variety of interests present.

The following is an approach to this emerging ecosystem. Specifically, it's an approach highlighting the underlying social relationships in these complex accountability processes and the results of our interviews and documentary review. We address three levels or layers of governance of algorithmic evaluations (macro, mezzo and micro), and explain their different components below.

Terminology matters

According to Bovens (2007), *algorithmic accountability* can be defined as the relationship between those who design or use algorithms and the forums that enforce the rules of conduct of the participating stakeholders. This involves certain requirements for action and results, holding stakeholders accountable and potentially subjecting them to consequences for their use of algorithms. These relationships can be assessed in different ways depending on the level of obligation (vertical, horizontal or diagonal) or the nature of the stakeholders (individual, collective, hierarchical or corporate).

Figure 1.
The three levels of governance in the evaluation of algorithms



Macro-level governance

Public, private and social spheres

The algorithmic evaluation ecosystem needs to account for the interaction between the public, private and social (or third) sectors. As some academic papers and white papers point out, the interdependence between the state and the market in the development and implementation of AI (Stahl et al. 2023) is crucial. While these exchanges may flow in a less systematic or informal way, the public sector needs to take the lead in these dynamics, as the business sector has difficulty self-regulating all the risks and impacts inherent to algorithms (Baeza-Yates and Matthews 2022).

In addition, the need to involve civil society in these dynamics arises from the expanding impact of algorithms in more areas of life, as well as the shared awareness that citizens need to have sufficient information to make well-informed decisions regarding their relationship with AI.

The balance in the relationship between the public and private sectors depends on the context. As noted, differing regional models of AI at a global level (North America, the European Union and China to start with) explain the different institutional arrangements and the economic, political, social and cultural traditions, etc. In some cases, the state has a greater influence and the development of regulations for the use and evaluation of AI focuses on its leadership; in others, market logic may influence how algorithm evaluations are regulated, based more on self-regulation and ongoing innovation in the AI field.

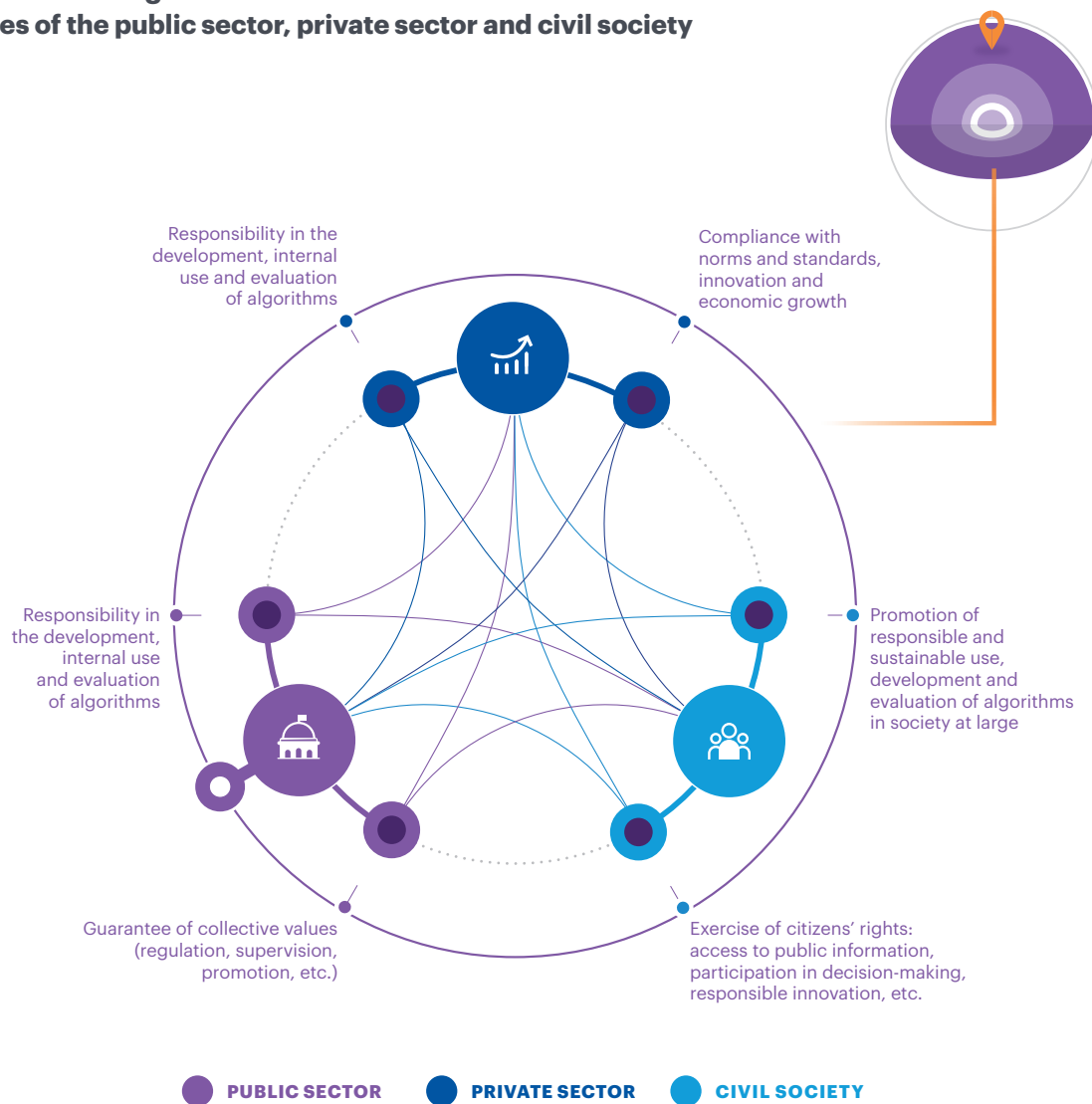
The role of civil society regarding AI evaluation is also significant insofar as it can serve as a benchmark for determining the balance in the relationship between states and markets and the position of certain values that determine collective coexistence, such as equality and freedom, etc.

Within this macro-level governance of algorithmic accountability, the role of the public sector is twofold: as user of algorithms and as guarantor of collective values, regardless of the position of the market and civil society in each geographic context. Various public interventions have already been launched to improve the implementation of this type of algorithm evaluation process. A recent paper (Basu et al. 2021) identified up to seven types of public actions (or policies) in relation to algorithmic accountability. Some of these coincide with the methods considered in this report. This is the case with principles and guidelines, impact assessments and implementations of technical and regulatory inspections.

Other cases, however, have a different scope. More than methods of evaluation, these are mechanisms designed by governments and administrations as a framework for their relationship with the private sector and civil society. These include prohibitions and moratoriums, the promotion of public transparency, the existence of independent external regulatory bodies, the right to be heard and appeal decisions and the conditions of AI procurement.

Here there is a direct relationship between the role of the private sector⁹ and civil society,¹⁰ so that it is necessary to design mechanisms that facilitate the governance of the algorithm evaluation process, integrating the three areas involved (the public and private sectors and civil society) as harmoniously as possible.

Figure 2.
Macro level of governance.
Roles of the public sector, private sector and civil society



Data source: own creation

⁹ For instance, in the conditions for public procurement of AI or prohibitions on companies to develop certain high-risk algorithms.

¹⁰ For example, the right to be heard and to appeal decisions as proposed by the EU General Data Protection Regulation, or the publicly accessible explanation of algorithm-based decisions, so that anyone can understand their content, regardless of their personal characteristics or level of formal education.

Mezzo-level governance

Activity sectors (health, education, transport, communications, banking, energy, security, etc.)

With the accelerated advances of AI, virtually any activity sector can use algorithms to automate processes to a greater or lesser extent. It is therefore necessary to articulate appropriate algorithmic accountability governance mechanisms within each of these intermediate (mezzo-level) activity sectors, as differentiated situations can occur.

Nevertheless, a certain degree of standardization is also required among sectors, from health, education, security, etc. (generally, closer to the public sector), to transport, energy, banking, telecommunications, etc. (generally closer to the private sector).

Algorithmic assessments should not be reserved for certain areas but should permeate the entire AI ecosystem focusing on the most wide-ranging effects as well as accounting for the other activity sectors they may have a relationship with. Recognizing this, and as highlighted by some of the interviewees in this project, sectors that use algorithms that can have a more direct impact on people's lives need special attention (such as health and financial services); however, other relevant criteria should also be taken into account, such as the implications to the environment, the effects on the rights of certain vulnerable groups or the effects on the evolution of each activity sector.

The proposed EU AI Act provides a roadmap for moving forward on this issue. While it includes general standards, it underscores some high-risk sectors, such as law enforcement, justice administration, immigration and border control, and access to basic services. In these priority areas algorithm evaluations are indispensable. In addition to general guidelines, specific standards for each sector may be needed to address the particularities of each,¹¹ though there is no consensus on this.

If this approach is followed, evaluation processes should be prioritized when algorithms are used in specific activities. For instance, the most invasive actions include people surveillance (such as facial recognition), profiling, classification of individuals and automated decision-making on the allocation of public services and social benefits. Some practices, such as real-time biometric monitoring, are prohibited in the proposed EU AI Act as they pose unacceptable risks. Looking at the ecosystem of evaluations holistically, all sectors must be aware of the possible impacts of the use of these systems for certain purposes and take the necessary measures to prevent and mitigate such problems.

¹¹ One example along these lines is the algorithmic impact assessment developed by the Ada Lovelace Institute for the UK health system (Groves 2022), which demonstrated the importance of adapting its processes to the specific characteristics of each area and geographic space.

Figure 3.
Mezzo level of governance.
Impact on different areas of activity



Data source: own creation

Micro-level governance

Stakeholders with a direct or indirect role (implementing organizations, external analysts, developers, users, etc.)

As noted, algorithmic evaluations can be classified according to those conducting the process (both organizations and people), but here we expand the focus to those who have some kind of relationship with it, whether direct or indirect. Costanza-Chok et al. (2022) identify three types of stakeholders with a direct role: internal staff of user organizations (first-party), consultants and other specialized organizations (second-party) and evaluation bodies or independent research staff (third-party). These three types form a key part of the algorithmic assessment ecosystem and play very distinct roles.

Stakeholders also exist that have an indirect role in the evaluation process, but they should be considered part of this micro level of governance. These include both users (and non-users) of the systems as well as other companies and organizations, regulatory or supervisory bodies, development companies, and the other organizations and civil society individuals involved.

Stakeholders with a direct role

Organizations that use algorithms (first-party) are central players in the algorithmic evaluation process. They have a key responsibility to ensure the availability of data on the technical and non-technical operation of algorithmic systems and the organizational dynamics that influence their deployment. This information is essential to safeguard genuine accountability and for the responsible and ethical development of algorithms. Along with companies in different sectors, the role of public administrations in their use of algorithms (in which users often don't have the choice to opt out) is particularly significant. Their responsibility is great because the areas of activity in which they operate have a direct impact on people's lives and consequences for their equality and freedoms.

Ultimately, the implementing organizations hold primary responsibility for ensuring that the functioning of the algorithms they use in their activities complies with current regulations and complies with the basic social, ethical and human rights standards that increasingly more international organizations are demanding.

Second, **consultancies and other organizations specializing in algorithmic evaluations (second-party)** provide an external viewpoint and the necessary knowledge to conduct the process appropriately, at the requirement of the implementing organizations, while meeting professional standards with certifications or seals of guarantee, for instance, that ensure the quality of the evaluations. Goodman et al. stress that without the appropriate standards and the necessary regulatory and control mechanisms, there is a risk that these types of specialized companies will not have sufficient freedoms to carry out their work independently and that the user organizations will perceive the process as an opportunity to clean up their image, which can be labelled *algorithm washing* or *algowashing*, particularly in cases that bring into question the existence of biases or breaches of the protection of personal data (Goodman and Trehu 2022; Schellmann 2021; persons interviewed).

In short, these stakeholders are key to carrying out algorithm evaluation tasks, especially from a technical perspective. However, the foreseeable future proliferation of algorithmic evaluations will require some type of inspection or control by public or private supervisory or standardization entities to professionalize the process, as well as standardize its deployment in different contexts to increase the confidence of business and society.

Third, **evaluation organizations or external research staff (third-party)** includes different independent algorithm monitoring initiatives, as well as academic staff, activists, journalists, etc. These stakeholders have the main advantage of being independent and doing their work without the need for a prior request from the organization evaluated. Certain issues also arise in these cases, such as the fact that they do not always have the resources or direct access to the data of the implementing organization to carry out their work (Costanza-Chok et al. 2022).

In addition, these actors may not feel bound to professional codes of conduct in their evaluations or may not explain their potential conflicts of interest. In short, it would be desirable for this stakeholder group to play an active role in algorithmic accountability processes, while respecting the autonomy of the implementing organizations. This could take the shape of agreements for data transfer in exchange for evaluation results or the promotion of periodic measurements as an opportunity for joint learning.

Stakeholders with an indirect role

In addition to the three stakeholder types mentioned, others, such as **users**, play an indirect role in the governance of algorithmic accountability. As Stahl et al. (2023) note, algorithmic evaluations must consider how the experiences and perceptions of the algorithm users are interrelated, to gain a comprehensive view of the risks and impacts to them, especially groups that may be particularly exposed.

The main objective of many of the methods mentioned above is to integrate system users into the algorithmic evaluation ecosystem, and even non-users (to carry out experiments and test hypotheses for instance). This is nothing more than a guarantee that there will be a social dimension to the evaluation process, as well as other more organizational or technical aspects, so that they can be conducted with a focus on people and better protect their rights.

Users are not only considered those outside the organizations, they are also the staff of the **organizations that implement algorithms** and that have direct contact with these systems, such as the staff of an organization using algorithm-based tools to make decisions regarding public services. In this case it is important to address the various aspects of the use of algorithms in a variety of contexts.

Algorithmic regulatory or supervisory entities play a crucial role not only at the initial stage of the process, but throughout the algorithm's life cycle. As noted, emphasis has been on the need to create public algorithm registers, as well as to promote the implementation of regulatory agencies or algorithmic supervisory bodies that contribute to the evaluation processes, among other things.

One of the most advanced recent such cases is the creation of the Spanish Agency for the Supervision of Artificial Intelligence,¹² the first of its kind in the EU, though similar initiatives exist in Canada,¹³ the United Kingdom¹⁴ and Singapore.¹⁵ Public oversight agencies will play an essential role in the algorithmic accountability ecosystem through direct and indirect methods of action, which will depend on the specific institutional and regulatory context, as well as on the priorities of each country.

Algorithm developers also play a critical role in understanding algorithmic assessment dynamics. These organizations are at the root of the process as they are responsible for creating the algorithms that will be implemented later by other companies and public administrations. Apart from the professional associations that may be created in future, these companies evidently have a responsibility in the algorithm evaluation process, as they must comply with the principles, ethical standards and technical regulations that are progressively proliferating in different institutional contexts. They are required to consider from the outset that their activities will be especially supervised, at the same time as they defend their copyright over the algorithms that generate or drive their capacity to develop future innovations in the field.

Finally, other **civil society organizations** or human rights and digital rights defenders, disability advocacy groups, media and journalists, jurists, university professors, etc., can also play a role in the process. Rather than having direct involvement, these stakeholders are called upon to collaborate with the algorithm evaluation ecosystem by disseminating public information, for instance, or raising social awareness of the impacts of the algorithms of private companies, or giving expert analysis in the media, academic spheres, etc.,.

In sum, the greater the transparency of algorithmic systems – the more data and evidence that is made available to the public – the greater the technical strength, institutional support and social trust they will gain. This will also increase social awareness of the new challenges and opportunities this technology entails.

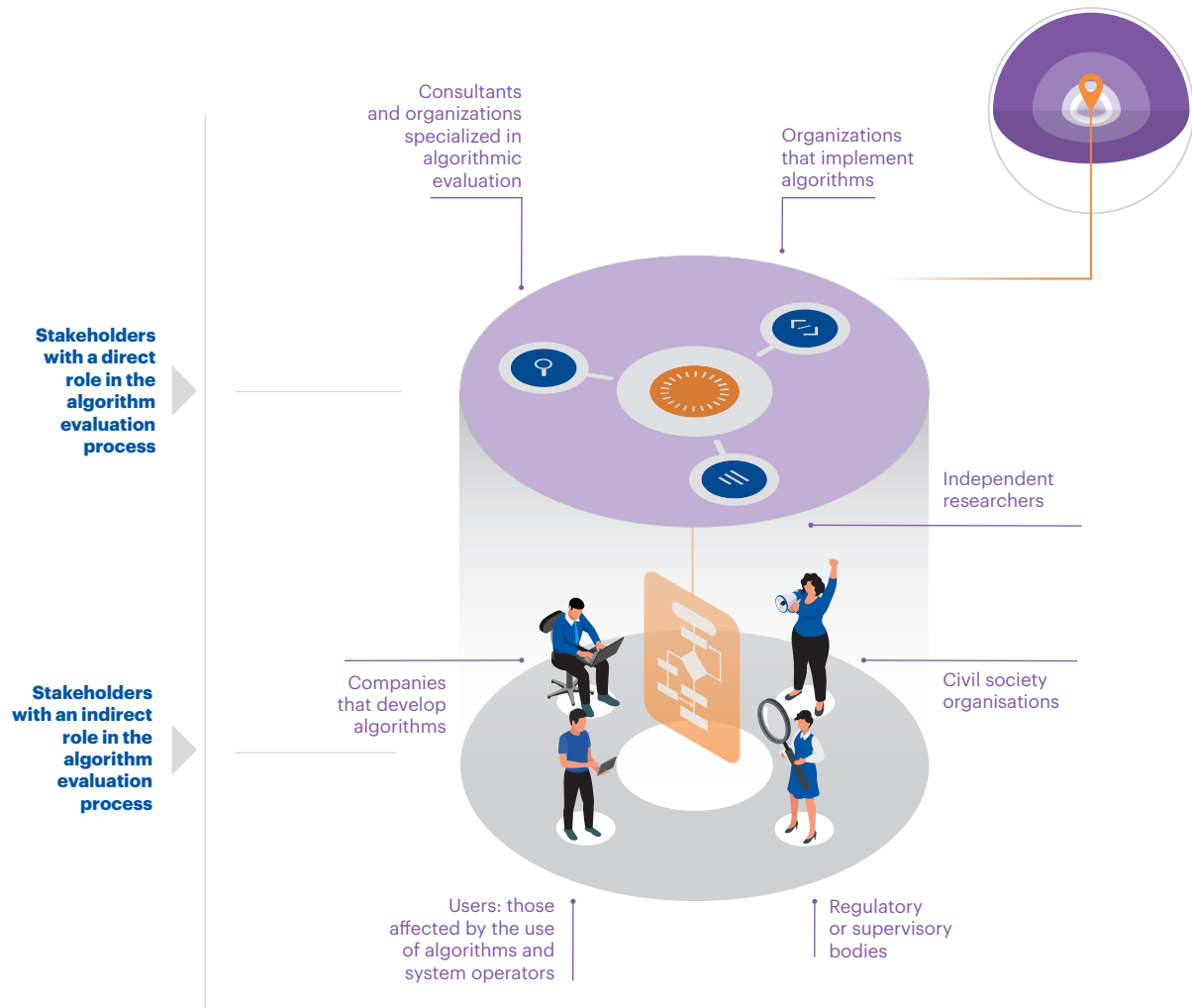
¹² This entity's statute was approved by Royal Decree 729/2023 of 22 August. Article 4 of its Annex states that the body "will have the functions of inspection, verification, sanction and other functions conferred upon it by the applicable European legislation and, in particular, in the field of AI".

¹³ Office of the Chief Information Officer (LEISURE), Treasury Board of Canada Secretariat (TBS) with its Algorithmic Impact Assessment tool: <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html>.

¹⁴ Centre for Data Ethics and Innovation and its Techniques for assuring AI systems: <https://cdeiuk.github.io/ai-assurance-guide/techniques/>.

¹⁵ Under the leadership of Infocomm Media Development Authority, one case of interest is the AI Verify Foundation (<https://aiverifyfoundation.sg>), which includes companies implementing AI, solution developers, and users or decision makers in the public sector.

Figure 4.
Micro level of governance.
Stakeholders with a direct or indirect role



Data source: own creation



5. Looking to the future: improving algorithmic evaluation processes

This document concludes with a section of practical points aimed at improving algorithm evaluation processes. It is based on the bibliographic and documentary review, interviews with experts, our proposal for dimensions for algorithmic evaluations, the applicable methods and the different levels of the algorithmic accountability ecosystem. From this information we propose six areas for improvement for the future of the evaluation process.

1. Explore new methods to develop algorithmic evaluations.

First, qualitative methods used to obtain details of the experiences and perceptions of algorithms from the staff of organizations and the public should be used more often. It would also be ideal to combine these methods with the more technical approaches generally used in these evaluations. Thus, algorithmic systems can be understood in all their complexity, and more holistic prevention and mitigation measures offered.

2. Create diverse and multidisciplinary algorithmic evaluation teams.

The variety of existing approaches and methods should drive the creation of multidisciplinary teams that combine technical knowledge with training in the social sciences. Depending on the focus of the evaluation and the sector, specialists in specific areas (e.g. human rights, health, transport, defence, etc.) enrich the process with their expert viewpoints. It is also important to drive socio-demographic and cultural diversity to incorporate different perspectives based on personal and collective experiences.

3. Strengthen the role of users and civil society organizations in understanding the impact of algorithms.

In line with the previous point, there is potential to involve society in general and the groups most affected by algorithmic decisions. Some of the tools explained in previous sections (such as IndieLabel) offer ideas on how to engage real users and potentially harmed communities in algorithmic evaluation processes, to critically reflect on the theoretical constructs and assumptions behind the algorithms and to understand the impact of these systems on certain groups. The entire evaluation process must be people oriented.

4. Promote social responsibility in the use of AI by the business sector.

Companies must prioritize responsible use of AI, specifically with the design and implementation of ethical and green algorithms. That is, they must incorporate aspects related to ethics and human rights, as well as environmental sustainability from the outset of the AI life cycle.¹⁶ The 2030 Agenda and its 17 Sustainable Development Goals (SDGs) offers a roadmap to achieve this goal.

The idea is to generate synergy among several areas of social responsibility. The algorithmic systems used should be aligned with this vision and the AI should serve as a vehicle to achieve objectives for society's benefit. Evaluations are essential to ensure that algorithms deployed by private companies follow these standards – always from a perspective of collaboration and support.

5. Strengthen public sector regulatory and supervisory work and its responsible implementation of algorithms.

In the varied ecosystem of algorithmic evaluations, the public sector must necessarily play an important role. While each context has its own dynamics, many experts agree that the public sector's role in defining the standards and regulating the processes is fundamental.

Public sector leadership should drive debate and encourage collaboration between the different sectors to create policies, guidelines and oversight bodies, among other measures, that promote balance between technological innovation and oversight, as well as good professional practices for algorithmic evaluations. There may also be a need for debate on public contracts, standardizations and changes to be made to pre-existing rules.

¹⁶ Spain's National Green Algorithms Plan points in this direction: https://portal.mineco.gob.es/RecursosNoticia/mineco/prensa/noticias/2022/20221213_plan_algoritmos_verdes.pdf

6. Promote the development of a mature algorithm evaluation ecosystem integrating levels and a global perspective.

The ecosystem of algorithmic evaluations is growing rapidly, but not necessarily in a coordinated way. Strengthening the relationships between various stakeholders and sectors is essential for defining appropriate standards and principles. Involving companies, governments, civil organizations, universities, the media and the public will encourage and enrich debate on pressing issues, such as transparency of the evaluations themselves, potential conflicts of interest and professional integrity in algorithmic evaluations.

In essence, contributing to the flourishing of an algorithmic evaluation ecosystem is pivotal in consolidating shared visions that transcend national and regional boundaries. The process should extend beyond governance limitations to encompass public, private and social sectors, different areas of activity and the diverse stakeholders involved. Harmonious integration of these elements is needed to collectively progress and improve the outcomes of algorithmic evaluation processes for the benefit of society.

Conclusions

This report is an effort to understand the **implications of algorithmic evaluations in a context marked by increasing use of AI in different areas of life**. It presents key concepts, tools and methods for developing algorithm evaluations incorporating different perspectives. It also addresses the ecosystem of the stakeholders and sectors involved with the aim of understanding the issue in all its complexity. Finally, it makes six proposals for improvement and to move towards algorithmic evaluations that are of value to contemporary societies.

In particular, the question that guides the research, and which has a clear practical focus, is the following: **How can algorithms be evaluated to detect any potential problems they contain and/or that may arise from their use, and how can these be mitigated?** The report stresses the importance of understanding the algorithmic evaluation process with a broad, holistic outlook to mitigate the risks and increase the benefits for the common good. To this end, it is necessary to combine purely technological knowledge with approaches from disciplines such as law, sociology, political science, psychology, philosophy, etc., in order to take all possible aspects into account when addressing the impact of algorithmic systems. Only then can we move towards responsible design and implementation of algorithms while respecting ethical principles and the rights of individuals and organizations.

To achieve this goal, we should aim for an algorithmic accountability ecosystem that **unites the efforts of public sector, private sector and third sector organizations**. There is a need for increased collaboration across different areas of action in which algorithms can have a significant impact. Users, both inside and outside organizations, should also play an important role in the evaluation process – especially groups affected by algorithmic decisions – as well as more vulnerable groups. A more open, participatory and collaborative approach to evaluations should lead to fundamental changes in how algorithms are designed and implemented, with the aim of ensuring that people remain at the forefront throughout the algorithm's life cycle.

References

- Ada Lovelace Institute. (2020). Examining the Black Box. Tools for assessing algorithmic systems. [PDF] Available at: <https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/> (Accessed: November 13, 2023).
- Baeza-Yates, R. and Matthews, J. (2022). Statement of Principles for Responsible Algorithmic Systems. ACM. [PDF] Available at: <https://www.acm.org/binaries/content/assets/public-policy/final-joint-ai-statement-update.pdf> (Accessed: December 16, 2023).
- Bandy, J. (2021). Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. Proceedings of the ACM on Human-Computer Interaction 5(CSCW1), 74:1–74:34. [PDF] Available at: <https://dl.acm.org/doi/10.1145/3449148> (Accessed: November 13, 2023).
- Basu, T., Brennan, J., Kak, A. and Joshi, D. (2021). Algorithmic accountability for the public sector. Learning from the first wave of policy implementation. Ada Lovelace Institute, AI Now and Open Government Partnership. [PDF] Available at: <https://www.opengovpartnership.org/wp-content/uploads/2021/08/executive-summary-algorithmic-accountability.pdf> (Accessed: November 13, 2023).
- Baykurt, B. (2022). Algorithmic accountability in U.S. cities: Transparency, impact, and political economy. Big Data & Society 9(2). [PDF] Available at: <https://journals.sagepub.com/doi/10.1177/20539517221115426> (Accessed: November 13, 2023).
- Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework. European Law Journal 13(4), 447–468. [PDF] Available at: <https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1468-0386.2007.00378.x> (Accessed: November 13, 2023).
- Brown, S., Davidovic, J. and Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. Big Data & Society 8(1). [PDF] Available at: <https://journals.sagepub.com/doi/10.1177/2053951720983865> (Accessed: November 13, 2023).
- Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research 81. [PDF] Available at: <https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf> (Accessed: November 13, 2023).
- Burr, C. and Leslie, D. (2022). Ethical Assurance: A Practical Approach to the Responsible Design, Development, and Deployment of Data-Driven Technologies. AI and Ethics 3(1), 1–26. [online] Available at: <https://link.springer.com/article/10.1007/s43681-022-00178-00> (Accessed: November 13, 2023).
- Costanza-Chock, S., Raji, I. D. and Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency 1571–1583. [PDF] Available at: <https://dl.acm.org/doi/10.1145/3531146.3533213> (Accessed: November 13, 2023).

- Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. [online] Available at: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G> (Accessed: November 13, 2023).
- De Manuel, A., Delgado, J., Parra Jounou, I., Ausín, T., Casacuberta, D., Cruz, M., Guersenzvaig, A., Moyano, C., Rodríguez-Arias, D., Rueda, J. and Puyol, A. (2023). Ethical assessments and mitigation strategies for biases in AI-systems used during the COVID-19 pandemic. *Big Data & Society* 10(1). [online] Available at: <https://journals.sagepub.com/doi/10.1177/20539517231179199> (Accessed: November 13, 2023).
- DeVito, M. A. (2017). From Editors to Algorithms. *Digital Journalism* 5(6), 753–773. [PDF] Available at: <https://www.tandfonline.com/doi/full/10.1080/21670811.2016.1178592> (Accessed: November 13, 2023).
- DeVos, A., Dhabalia, A., Shen, H., Holstein, K. and Eslami, M. (2022). Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. *CHI Conference on Human Factors in Computing Systems* 1–19. [PDF] Available at: <https://dl.acm.org/doi/pdf/10.1145/3491102.3517441> (Accessed: November 13, 2023).
- Dodge, J., Khanna, R., Irvine, J., Lam, K., Mai, T., Lin, Z., Kiddle, N., Newman, E., Anderson, A., Raja, S., Matthews, C., Perdriau, C., Burnett, M. and Fern, A. (2021). After-Action Review for AI (AAR/AI). *ACM Transactions on Interactive Intelligent Systems* 11(3–4), 29:1–29:35. [PDF] Available at: <https://dl.acm.org/doi/pdf/10.1145/3453173> (Accessed: November 13, 2023).
- Eriksson, M. and Johansson, A. (2017). Tracking Gendered Streams. *Culture Unbound*, 9(2), 163–183. [PDF] Available at: <https://www.diva-portal.org/smash/get/diva2:1243114/FULLTEXT01.pdf> (Accessed: November 13, 2023).
- Eticas Consulting. (n.d.). ¿Cómo se audita un algoritmo? Los cinco pasos de una auditoría algorítmica. (How do you audit an algorithm? The five steps of an algorithmic audit.) [online] Available at: <https://www.eticasconsulting.com/como-se-audita-un-algoritmo-pasos-para-auditar-algoritmos/> (Accessed: November 13, 2023).
- Eubanks, V. (2019). *La automatización de la desigualdad. Herramientas de tecnología avanzada para supervisar y castigar a los pobres.* (The automation of inequality. Advanced technology tools to monitor and punish the poor.) Madrid, Spain: Capitán Swing.
- Galdon Clavell, G., Martín Zamorano, M., Castillo, C., Smith, O. and Matic, A. (2020). Auditing Algorithms: On Lessons Learned and the Risks of Data Minimization. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 265–271. [PDF] Available at: <https://dl.acm.org/doi/pdf/10.1145/3375627.3375852> (Accessed: November 13, 2023).
- Garde Roca, J. A. (2023). ¿Pueden los algoritmos ser evaluados con rigor? (Can algorithms be rigorously evaluated?) *Encuentros Multidisciplinares, (Can algorithms be rigorously evaluated? Multidisciplinary Encounters)* 73, 1–13. [PDF] Available at: <http://www.encuentros-multidisciplinares.org/revista-73/juan-antonio-garde.pdf> (Accessed: November 13, 2023).
- Godin, K., Stapleton, J., Kirkpatrick, S. I., Hanning, R. M. and Leatherdale, S. T. (2015). Applying systematic review search methods to the grey literature: a case study examining guidelines for school-based breakfast programs in Canada. *Systematic Reviews* 4(1), 138. [PDF] Available at: <https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-015-0125-0> (Accessed: November 13, 2023).

- Goodman, E. P. and Trehu, J. (2022). AI Audit-Washing and Accountability. GMF. [PDF] Available at: <https://www.gmfus.org/news/ai-audit-washing-and-accountability> (Accessed: November 13, 2023).
- Groves, L. (2022). Algorithmic impact assessment: a case study in healthcare. Ada Lovelace Institute. [PDF] Available at: <https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/> (Accessed: November 13, 2023).
- Hamilton, M. (2021). Evaluating Algorithmic Risk Assessment. *New Criminal Law Review* 24(2), 156–211. [PDF] Available at: <https://online.ucpress.edu/nclr/article/24/2/156/116809/Evaluating-Algorithmic-Risk-Assessment> (Accessed: November 13, 2023).
- Kelly-Lyth, A. and Thomas, A. (2023). Algorithmic management: Assessing the impacts of AI at work. *European Labour Law Journal* 14(2), 230-252. [PDF] Available at: <https://journals.sagepub.com/doi/10.1177/20319525231167478> (Accessed: November 13, 2023).
- Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., Ahamat, G., Leutner, F., Goebel, R., Knight, A., Adams, J., Hitrova, C., Barnett, J., Nachev, P., Barber, D., Chamorro-Premuzic, T., Klemmer, Gregorovic, M., Khan, S. and Lomas, E. (2021). Towards Algorithm Auditing. A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms. *SSRN Electronic Journal*. [PDF] Available at: <https://discovery.ucl.ac.uk/id/eprint/10164738/1/owards%20Algorithm%20Auditing%20A%20Survey%20on%20Managing%20Legal,%20Ethical%20and%20Technological%20Risks%20of%20AI,%20ML%20and%20Associated%20Algorithms.pdf> (Accessed: November 13, 2023).
- Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P. and Karahalios, K. (2019). Investigating political bias in social media and web search. *Information Retrieval Journal* 22(1), 188–227. [PDF] Available at: <https://link.springer.com/article/10.1007/s10791-018-9341-2> (Accessed: November 13, 2023).
- Lam, M. S., Pandit, A., Kalicki, C. H., Gupta, R., Sahoo, P. and Metaxa, D. (2023). Sociotechnical Audits: Broadening the Algorithm Auditing Lens to Investigate Targeted Advertising. *Proceedings of the ACM Human-Computer Interaction* [PDF] Available at: https://hci.stanford.edu/publications/2023/Lam_STA_CSCW23.pdf (Accessed: November 13, 2023).
- Mantelero, A. (2018). AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review* 34(4), 754–772. [online] Available at: <https://www.sciencedirect.com/science/article/pii/S0267364918302012?via%3Dihub> (Accessed: November 13, 2023).
- Meßmer, A.-K. and Degeling, M. (2023). Auditing Recommender Systems. Putting the DSA into practice with a risk-scenario-based approach. *Stiftung Neue Verantwortung*. [PDF] Available at: <https://www.stiftung-nv.de/de/publication/auditing-recommender-systems> (Accessed: November 13, 2023).
- Metcalf et al. (2021). Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. *FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. [PDF] Available at: <https://dl.acm.org/doi/pdf/10.1145/3442188.3445935> (Accessed: November 13, 2023).

- Minkinen, M., Laine, J. and Mäntymäki, M. (2022). Continuous Auditing of Artificial Intelligence: A Conceptualization and Assessment of Tools and Frameworks. *Digital Society* 1(3), 21. [PDF] Available at: <https://link.springer.com/article/10.1007/s44206-022-00022-2> (Accessed: November 13, 2023).
- Mökander, J., Axente, M., Casolari, F. and Floridi, L. (2022). Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds and Machines* 32(2), 241–268. [PDF] Available at: <https://link.springer.com/article/10.1007/s11023-021-09577-4> (Accessed: November 13, 2023).
- Morondo, D. and Eguiluz, J. A. (2022). La discriminación algorítmica en España: límites y potencial del marco legal. (Algorithmic discrimination in Spain: limits and potential of the legal framework.) Digital Future Society Think Tank. [PDF] Available at: <https://digitalfuturesociety.com/es/report/algorithmic-discrimination-in-spain/> (Accessed: November 13, 2023).
- Munz, P., Hennick, M. and Stewart, J. (2023). Maximizing AI reliability through anticipatory thinking and model risk audits. *AI Magazine* 44(2), 173–184. [PDF] Available at: <https://onlinelibrary.wiley.com/doi/10.1002/aaai.12099> (Accessed: November 13, 2023).
- Novelli, C., Casolari, F., Rotolo, A., Taddeo, M. and Floridi, L. (2023). Taking AI risks seriously: A new assessment model for the AI Act. *AI & SOCIETY*. [PDF] Available at: <https://link.springer.com/article/10.1007/s00146-023-01723-z> (Accessed: November 13, 2023).
- Oswald, M., Grace, J., Urwin, S. and Barnes, G. C. (2018). Algorithmic risk assessment policing models: Lessons from the Durham HART model and ‘Experimental’ proportionality. *Information & Communications Technology Law* 27(2), 223–250. [PDF] Available at: <https://www.tandfonline.com/doi/epdf/10.1080/13600834.2018.1458455?needAccess=true> (Accessed: November 13, 2023).
- Papakyriakopoulos, O. and Mboya, A. M. (2023). Beyond Algorithmic Bias: A Socio-Computational Interrogation of the Google Search by Image Algorithm. *Social Science Computer Review* 41(4), 1100–1125. [PDF] Available at: <https://journals.sagepub.com/doi/10.1177/08944393211073169> (Accessed: November 13, 2023).
- Pappu, A., Brennan, J., Strait, A., Parker, I. and Jones, E. (2021). Technical methods for regulatory inspection of algorithmic systems in social media platforms (Ethics and accountability in practice). Ada Lovelace Institute. [PDF] Available at: https://www.adalovelaceinstitute.org/wp-content/uploads/2021/12/ADA_Technical-methods-regulatory-inspection_report.pdf (Accessed: November 13, 2023).
- Percy, C., Dragicevic, S., Sarkar, S. and d’Avila Garcez, A. (2021). Accountability in AI: From principles to industry-specific accreditation. *AI Communications* 34(3), 181–196. [PDF] Available at: <https://content.iospress.com/articles/ai-communications/aic210080> (Accessed: November 13, 2023).
- Raji, I. D. and Buolamwini, J. (2019). Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* 429–435. [PDF] Available at: <https://dl.acm.org/doi/pdf/10.1145/3306618.3314244> (Accessed: November 13, 2023).
- Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T. and Ghani, R. (2019). Aequitas: A Bias and Fairness Audit Toolkit. *arXiv*. [PDF] Available at: <https://arxiv.org/abs/1811.05577> (Accessed: November 13, 2023).
- Sandu, I., Wiersma, M. and Manichand, D. (2022). Time to audit your AI algorithms. *Maandblad Voor Accountancy En Bedrijfseconomie* 96(7/8). [PDF] Available at: <https://mab-online.nl/article/90108/> (Accessed: November 13, 2023).

Sandvig, C., Hamilton, K., Karahalios, K. and Langbort, C. (2014). Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. University of Michigan. [PDF] Available at: <https://websites.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf> (Accessed: November 13, 2023).

Schellmann, H. (2021). Auditors are testing hiring algorithms for bias, but there's no easy fix. MIT Technology Review. [Online] Available at: <https://www.technologyreview.com/2021/02/11/1017955/auditors-testing-ai-hiring-algorithms-bias-big-questions-remain/> (Accessed: November 13, 2023).

Sheehan, M., & Du, S. (2022). What China's Algorithm Registry Reveals about AI Governance. Carnegie Endowment for International Peace. [Online] Available at: <https://carnegieendowment.org/2022/12/09/what-china-s-algorithm-registry-reveals-about-ai-governance-pub-88606> (Accessed: October 3, 2023).

Sloane, M. (2021). The Algorithmic Auditing Trap. OneZero. [Online] Available at: <https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d> (Accessed: 13-11-2023).

Spyridou, P. (Lia), Djouvas, C. and Milioni, D. (2022). Spyr Modeling and Validating a News Recommender Algorithm in a Mainstream Medium-Sized News Organization: An Experimental Approach. Future Internet 14(10) [PDF] Available at: <https://www.mdpi.com/1999-5903/14/10/284> (Accessed: November 13, 2023).

Srba, I., Moro, R., Tomlein, M., Pecher, B., Simko, J., Stefancova, E., Kompan, M., Hrcckova, A., Podrouzek, J., Gavornik, A. and Bielikova, M. (2023). Auditing YouTube's Recommendation Algorithm for Misinformation Filter Bubbles. ACM Transactions on Recommender Systems 1(1). [online] Available at: <https://dl.acm.org/doi/10.1145/3568392> (Accessed: November 13, 2023).

Stahl, B. C., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B., Kirichenko, A., Marchal, S., Rodrigues, R., Santiago, N., Warso, Z. and Wright, D. (2023). A systematic review of artificial intelligence impact assessments. Artificial Intelligence Review 56(11). [PDF] Available at: <https://link.springer.com/article/10.1007/s10462-023-10420-8> (Accessed: November 13, 2023).

Stanford University Human-Centered Artificial Intelligence (2023). Artificial Intelligence Index Report 2023. [PDF] Available at: https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (Accessed: November 13, 2023).

Strubell, E., Ganesh, A. and McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. arXiv. [PDF] Available at: <https://arxiv.org/abs/1906.02243> (Accessed: November 13, 2023).

Wasilow, S. and Thorpe, J. B. (2019). Artificial Intelligence, Robotics, Ethics, and the Military: A Canadian Perspective. AI Magazine 40(1), 37–48. [PDF] Available at: <https://dl.acm.org/doi/pdf/10.1145/3306618.3314244> (Accessed: November 13, 2023).

Wilson, C., Mislove, A., Ghosh, A. and Jiang, S. (2020). Auditing the pymetrics Model Generation Process. [PDF] Available at: https://cbw.sh/static/audit/pymetrics/pymetrics_audit_result_whitepaper.pdf (Accessed: November 13, 2023).

Yam, J. and Skorburg, J. A. (2021). From human resources to human rights: Impact assessments for hiring algorithms. Ethics and Information Technology 23(4), 611–623. [online] Available at: <https://link.springer.com/article/10.1007/s10676-021-09599-7> (Accessed: November 13, 2023).

Appendix

Methodology

The preparation of this report involved field work with three main phases: a) a systematic review of the academic literature, b) documentary analysis of the grey literature, and c) in-depth interviews with specialists and key stakeholders in algorithmic evaluation processes, in Spain, United Kingdom, Holand and the United States. We explain the most significant details of each stage below.

The research question that guided the review of both the academic literature and grey literature was as follows: What are the main tools and methodologies for algorithm evaluations found in the literature? The main aim of these stages of the field research was to identify and systematize the information published on the subject to date.

Systematic review of the academic literature

In this stage we reviewed scientific articles published in journals indexed in the Journal Citation Reports (JCR) and SCImago Journal Rank (SJR) databases, as well as relevant conference papers. The search was carried out in the Web of Science database in July 2023, with the following sequence of terms: TI = ("algorithm*" OR "AI" OR "artificial intelligence" OR "automated system*" OR "machine learning" OR "deep learning") AND TS = ("audit*" OR "assessment*"). In other words, we considered articles that included AI-related terms in the title and a subject related to evaluations.

We included all studies in the database from 1985 to 2023 and in the following areas of research: computer science (specifically, information systems and AI), information sciences, management, economics, the social sciences, business, law, sociology and public administration. This search yielded a total of 2,907 documents.

Once the results were obtained, the database with the relevant information for each article was downloaded and the inclusion criteria defined. Only articles with a focus on algorithmic evaluations or that included in their methodology information relevant to algorithmic evaluation methods, regardless of the sector, were considered in the final analysis.

Subsequently, we reviewed the titles and abstracts to select the most relevant texts for the final synthesis. For this process we used ASReview (<https://asreview.nl/>), an active learning and open-source tool used to streamline the systematic literature review process. By using it, the researcher labels articles as relevant or irrelevant, and the model is progressively trained to identify and prioritize the articles of greatest interest. As such it is not necessary to manually review the entire database to find the texts needed.

This stage unearthed 62 relevant articles. To compensate for possible failures of the method, the process was complemented with a manual web search for papers and recommendations from the experts we interviewed. This increased the database count to 90 documents. In the more detailed review of the articles, 26 were excluded as they did not meet the specific search and analysis criteria. The final number of articles included was 64.

Systematic review of reports and other publications

A search and analysis of the grey literature was also conducted to round out the information obtained in the academic literature review. This included reports and other types of publications (blog posts, websites, etc.) from public bodies and third sector organizations, universities, think tanks, companies and other entities. A conventional Google search was conducted in line with guidelines of the method explained by Godin et al. (2015), using the following sequences of terms:

- audit + algorithm
- audit + artificial intelligence
- assessment + algorithm
- assessment + artificial intelligence
- audit + algorithm
- algorithmic audit
- algorithm evaluation
- algorithmic evaluation

The first 100 results of each term sequence were reviewed. Specifically, a quick reading of each title and summary to verify whether the documents specifically addressed the issue of audits and algorithm evaluations. If so, they were included in the database for subsequent data extraction and the synthesis of results.

To prevent the loss of relevant results to the extent possible and in line with previous studies (Godin et al. 2015), this search was complemented with a manual review of the web pages of some AI and algorithm bodies of reference such as Ada Lovelace Institute, AI Now Institute, AI Watch, Stanford University Human-Centered Artificial Intelligence, OECD and other European bodies, etc. Table 4 details the number of documents included in the review.

Table 4.
Grey literature documents for each type reviewed

Type of document	Number of documents
Official documents	11
Reports	18
Blog posts and publications on websites	23
Books and book chapters	3
Normative texts	5
Total	60

Data source: authors' data

Expert interviews

The third stage of the field research took place between August and September 2023. It consisted of 15 semi-structured interviews with specialists in AI and algorithmic evaluations, working in international organizations, European bodies, private companies and consultants, universities and third sector organizations (see Table 4). All interviews were conducted online using Google Meet or Zoom, ten in English and five in Spanish.

Table 5.
Persons interviewed according to the agency or organization of origin

Type of organization	Number of people
Independent investigative bodies	1
Third sector organizations	3
Private companies and consultants	2
European or international bodies	4
Academic institutions	3
Independent researchers	2
Total	15

Data source: authors' data

The purpose of the interviews was to obtain conceptual and practical information on the development of algorithmic evaluations in different contexts to corroborate the data obtained from the systematic literature review. To this end, we used the following 14 questions, which varied according to the dynamics of the discussion.

Starting questionnaire

- 1.** Please tell us your name and current position in your agency or organization.
- 2.** What are your current responsibilities? How do they relate to artificial intelligence in general and algorithm evaluations in particular?
- 3.** What do you think is the best definition of an algorithmic audit?
- 4.** What are the main types of audits and algorithmic evaluations? And what are the differences and similarities?
- 5.** What methodologies and tools for algorithm audits and evaluations you have experience with? What are their main characteristics? Could you give a specific example?
- 6.** Could you describe the process followed by your organization to conduct audits and algorithm evaluations? That is, who participates, how do you define strategies, communicate results, etc.?
- 7.** What is the legal and institutional framework defined in your context for conducting algorithm audits and evaluations?
- 8.** How do you think the public sector should conduct algorithm audits and evaluations?
- 9.** Are you aware of any cases of public administrations currently carrying out this type of audit?
- 10.** In your experience, what are the main lessons learned regarding algorithm audits and evaluations, that is, challenges, opportunities and the like?
- 11.** What do you think should be done to improve these processes in the future?
- 12.** Would you like to add any comments?
- 13.** Do you have access to a report or document that may provide information relevant to our research?
- 14.** In your opinion, should we consider another key person who you think it might be possible to interview on this topic?

Acknowledgements

Authors

J. Ignacio Criado is an associate professor in the Department of Political Science and International Relations, and Director of the Lab Innovación Tecnología y Gestión Pública (Innovation, Technology and Public Management Lab) (IT_GesPub) at the Autonomous University of Madrid. His interests focus on open government, digital government, public innovation, social media, as well as algorithmic governance and artificial intelligence in the public sector.

Ariana Guevara-Gomez is in the Department of Political Science and International Relations, and a researcher at the Innovation, Technology and Public Management Lab at the Autonomous University of Madrid. She researches gender, technology and artificial intelligence, innovation and public administration.

Research coordination

J. Ignacio Criado

Project coordination

Tanya Álvarez leads the Digital Future Society Think Tank research on digital divides and digitalisation of the public sector. She advocates for an interdisciplinary perspective of how technology impacts society. She has a degree in art history from Swarthmore College and a master's degree in cultural heritage management from the University of Barcelona.

Editing and design

Marta Campo, editor and proofreader

Manuela Mouliau, designer and author of infographics

Lara Cummings, English edition, translator and copyeditor

Interviewees

Adriano Soares Koshiyama, co-founder of the company Holistic AI

Albert Sabater, Serra Hunter Associate Professor of Sociology at the University of Girona and Director of the Observatory for Ethics in Artificial Intelligence of Catalonia

Aparna Surendra, Strategic Research and Insight Team manager at AWO Agency (based in London, Brussels and Paris)

Carlos Castillo, ICREA research professor and leader of the Web Science and Social Computing research group both at Pompeu Fabra University

Dafna Feinholz, Chief of the Bioethics and Ethics of Science Section at UNESCO

Fabio Curi, back-end developer and AI consultant for the OECD

Iban Garcia del Blanco, Member of the European Parliament

Javier de la Cueva, patron of the Civio Foundation and specialist in Law and Information Technology and Communication

Jurriaan Parie, member of the Algorithm Audit Board

Krishnaram Kenthapadi, Chief AI Officer and Chief Scientist of Fiddler AI

Lara Groves, researcher at the Ada Lovelace Institute in the UK

Ola Al Khatib, pre-doctoral researcher at Utrecht University and member of the Algorithm Audit team

Rashad Abelson, Technology Sector Lead and Due Diligence Legal Expert at the OECD

Ricardo Baeza-Yates, research director at the Institute for Experiential AI at Northeastern University, United States

Shazade Jameson, independent AI governance researcher and consultant, specifically on urban studies

Please cite this report as:

Digital Future Society (2024). Towards accountable algorithms – tools and methods for responsible use.

Contact details:

thinktank@digitalfuturesociety.com

Un programa de



GOBIERNO DE ESPAÑA

MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA

SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

red.es



Mobile WorldCapital Barcelona