

A Human Rights Reboot for the Global Digital Compact

Avaaz Foundation submission to the Global Digital Compact April 2023

Imagine a world where technology abides by human rules and human legislation, from government policies to human rights. That's the future I want to be in. A future where we leverage the power of technology but in a secure manner for us and our world.

– Sara, Avaaz member from Portugal

Contents

Introductio	on	1
Section 1	Thematic area 3 - Apply human rights online	2
Section 2	Thematic area 4 - Accountability criteria for discrimination and misleading	
content		5
Section 3	Thematic area 5 - Regulating AI	9
Annex I	Reporting and transparency to combat disinformation during operation	15
Annex II	Examples of questions that could be included in an AI HRIA questionnaire	17
Annex III	An AI Audit proposal to identify Moments of Bias	18

Introduction

Avaaz is a movement made up of tens of millions of people that is a dedicated defender of human rights, freedom of expression, and the principle that the Internet is for everyone. Avaaz simply would not exist without a free and open Internet. However, our research and engagement with Avaaz members, victims' representatives, civil society, and government agencies shows that the free and open digital future that benefits humanity so much is at risk.

The Digital Compact must stand at the heart of the world's determination to put humans first, before profit, before convenience, and before the competitive race to be first on the market with a new technological product. Human rights are historically-established standards and, along with the promotion of democracy, the rule of law, and the protection of our environment, human rights are the clear choice for the basis on which we should build our online futures. The Digital Compact should work to safeguard human rights, protect democracy, and care for our well-being.

Section 1 Apply human rights online

Please remember, AI and the needs of business are not more important than human rights. Let's use AI to improve the quality of life on this planet rather than leave it to the whims of corporations. The decisions we make now will shape the lives of generations to come. –Helen, Avaaz member from Belgium

Principle : "Human rights know no bounds - online or offline, they demand equal respect and protection."

Every person, no matter where they are or how they communicate, deserves to have their fundamental rights and freedoms upheld and protected. The online world is no exception. We cannot stand by and watch as digital spaces become breeding grounds for oppression, discrimination, and abuse. It is our duty to demand that online human rights receive the same level of respect and protection as those in the physical world. We must strive for a world where every individual's dignity and worth are recognized, both online and offline and not downgraded in the pursuit of capital.

The rights impacted by the digital sphere extend across all areas of rights protected in the real word. The right to free speech is undermined through disinformation, increasingly produced by AI software, and spread through bots and also as a result of the non transparent removal of online content by online service providers. Freedom of thought is undermined by online curation algorithms targeting content on the basis of opaque data use and the use of AI to interpret the emotional responses of service users. The right to privacy: is impacted by opaque data collection and online surveillance through data tracking and in the real world through real time, and recorded, surveillance of movements through cameras, especially those incorporating bio recognition software.

These rights are of course crucial, but other rights are also deeply impacted by the transfer of so much of our lives to the digital sphere and we have brought the voices of real people who need the Digital Compact to do more than it has managed to date in protecting their rights. Below are just a few examples of impacts on human rights beyond free speech and privacy:

The rights of the child:

In all actions concerning children, whether undertaken by public or private social welfare institutions, courts of law, administrative authorities or legislative bodies, the best interests of the child shall be a primary consideration (CRC, Art. 3.1).

States Parties undertake to ensure the child such protection and care as is necessary for his or her well-being, taking into account the rights and duties of his or her parents, legal guardians, or other

individuals legally responsible for him or her, and, to this end, shall take all appropriate legislative and administrative measures (CRC, Art. 3.2).

'It's horrific. It's just the bleakest of worlds. It's a world I don't recognise. It's a ghetto of the online world that once you fall into it, the algorithm means you can't escape it and keeps recommending more content, you can't escape it. – Ian Russell, father of Molly Russell

The tragic outcome of the lack of human rights consideration in the design and deployment of online recommender algorithms can be seen in the story of UK teenager Molly Russell. Molly was just 14 when she took her own life. After Molly died in 2017, her family looked into her Instagram account and found bleak depressive material, graphic self-harm content and suicide-encouraging memes. Her father believes this social media content encouraged her desperate state. He described the process clearly: "Online, Molly found a world that grew in importance to her and its escalating dominance isolated her from the real world. The pushy algorithms of social media helped ensure Molly increasingly connected to her digital life while encouraging her to hide her problems from those of us around her, those who could help Molly find the professional care she needed."

To apply the rights Molly should have enjoyed as a child, we urge the UN to adopt in full the principles and commitments to ensure accountability for discrimination and misleading content as described in Section 2 of this document.

The rights to life and to health:

Everyone has the right to life, liberty and security of person (UDHR, Art. 3).

Everyone has the right to a standard of living adequate for the health and well-being of himself and of his family... (UDHR, Art. 25).

"He would still be here." - Widow of suicide victim

The tragic words of a Belgian widow, who blames her husband's suicide on his relationship with an online generative AI chatbot that became his confidant and encouraged him to kill himself, show how urgently we need to assess AI's risks before it is released into the world. Generative AI, which can produce seemingly human responses, has been released at top speed in a race to corner the market, without pausing to assess the impact these ostensibly empathetic programmes can have. Even though many of these AI chatbots have only been with us for a few weeks, severe mental health issues are already emerging from their use. The apparent humanity of the AI – presenting themselves as emotive, giving responses as if they were talking to a friend, and establishing a bond and dependency – can lead to tragic results when interacting with vulnerable adults.

To apply the rights this vulnerable individual should have been able to rely on, we urge the UN to adopt in full the principles and commitments to regulate AI as described in Section 3 of this document.

The right to non-discrimination

All persons are equal before the law and are entitled without any discrimination to the equal protection of the law. In this respect, the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status (ICCPR, Art. 26).

Hate speech and misinformation fester on the platform - from posts glorifying Nazi attacks on trans rights and organisations, to slurs that trans women are paedophiles and where anti trans narratives slip past Facebook's outdated moderation policies, which fail to recognise common forms of transphobic content.

-Trans community representative whom Avaaz worked with on an investigation into anti trans hate on Facebook in 2021

Online algorithms can replicate existing discrimination, such as discrimination based on gender, disability, and race, and can amplify hate and polarise societies. One group identified by the UN as especially vulnerable to hate speech are the over 1.9 million people who were excluded from India's National Register of Citizens (NRC), published in the northeastern state of Assam on 31 August, 2019. <u>An Avaaz report</u> found that, rather than promoting community healing, social media propagated discrimination. Bengalis, Muslims in particular, faced an extraordinary chorus of abuse and hate in Assam on Facebook.

The report also exposed the limitations of Facebook's artificial intelligence (AI) driven strategy to detect hate speech and we see this entrenchment of discrimination <u>manifesting</u> <u>across AI technology</u>. For example, Facebook's algorithm learned only <u>to show ads for</u> <u>housing to white people</u>, as that generated better engagement statistics than showing them to minority groups.

Al has been reported to have <u>reduced online employability rankings for women</u> with young children in Croatia, whereas for men the parameter was not even displayed or taken into account. In a study of Spain's Al offender prediction software, predictions were found to be <u>biased against non-Spanish 'foreigners'</u>. People classified as Maghrebi, Latin American, European and Other were almost twice as likely to be wrongly classified as high-risk by the machine-learning models than Spanish nationals, and were less likely to be classed as unlikely to reoffend.

The right to just and favourable conditions of work

Everyone has the right to work, to free choice of employment, to just and favourable conditions of work and to protection against unemployment. Everyone, without any discrimination, has the right to equal pay for equal work (UDHR, Art. 23).

"You are a number, whether you like it or not, You are a number, you are an ID, you are a scooter icon that moves around a map. ... The app does not know you. The app does not listen to you." – Singh, Scooter courier from Mumbai

Workers like Singh, whose working conditions and job allocation are managed by AI, often called gig workers, face <u>lack of recognition</u> as employees and so are barred from the protections their societies offer to employees. This, combined with poor working conditions, such as a lack of stable or predictable working hours, <u>exacerbates already low levels</u> of pay. Other concerns include the <u>stress experienced</u> by moderation tech workers, who have to address huge volumes of distressing content to moderate social media, and the sweatshop labour conditions of those conforming data for large generative AI systems. All of these aspects of workers rights, which have been well documented, must form part of the Global Digital Compact's efforts to apply human rights online.

In conclusion:

These examples underline the need for understanding that human rights should underpin all the approaches in the Global Digital Compact - and our remaining responses will examine the implementation of this principle in two specific thematic areas: accountability criteria for discrimination and misleading content, and the regulation of AI.

Section 2: Accountability criteria for discrimination and misleading content

I write to you humbly as a concerned grandmother. I have two beautiful grandsons who mean the world to me. I am aware that nowadays, young people live, breathe and dream through social media. There's no harm in that. However, I am very much afraid of the harm social media can do. We are all aware of the misinformation bandied about and how vulnerable young people are. –Cath, Avaaz member from Malta

Avaaz's research on the damage disinformation can do to democracy and human lives dates back to 2019. A full report can be found here: <u>https://secure.avaaz.org/campaign/en/disinfo_hub/</u>. We've uncovered disinformation

operations designed to undermine confidence in the leadership in France, confuse climate consensus, smear political figures in the United States, and target vulnerable minorities in India.

Disinformation fosters division, often targeting vulnerable groups like refugees or minorities in violation of the right to non-discrimination. It increases polarisation, fracturing any ability to engage in productive civic discourse in ways that undermine freedom of thought.

Disinformation has a death toll. Doctors and nurses had to waste precious time speaking out against the infodemic of lies surrounding COVID-19 -- lies that kept some people at home during the pandemic who should have been treated and prompted others to take fake 'miracle cures,' like bleach or disinfectant.

This revelation is not new for us. Avaaz has been at the forefront of the public push to force social media platforms to deal with harmful content for almost four years. Avaaz has been one of the leading organisations in the EU, in the US, and around the world, reporting on disinformation and its scale, commissioning polls on its harmful effects, evaluating platforms' efforts to manage it, and identifying their failures. We have analysed platforms including Facebook, Youtube and WhatsApp. We are also one of the few NGOs currently co-drafting the new EU Code of Practice on Disinformation.

Principles:

Our accountability criteria for misinformation and hate speech online consists of four fundamental principles: ¹

- 1. **Safer algorithms** to make sure online services really take responsibility for the societal harms their services create, the focus must be on a systemic reform of how algorithmic systems operate. This will protect free speech and remove the structural drivers of harmful but legal content and discrimination.
- 2. **Unprecedented transparency** over access to data to allow society to finally understand the real scale of the problem and the effectiveness of the solutions; over what actions the platforms take with speech; and to help users find reliable sources of information to counter disinformation
- 3. Users' data belongs to them give back control and offer meaningful choices over what data is shared with services and the way algorithms use it to determine what we see and experience in our digital lives, especially where that use of data targets and amplifies harmful content.
- 4. Effective enforcement to be sure these commitments stick.

Commitments to achieve these principles include:

¹ We address the issue of discrimination in terms of the data sets and operation of AI in section 3.

- 1. Safer algorithms online platforms and service providers should:
 - a. Detoxify the algorithms of the platforms
 - i. **remove structural drivers of harmful but legal content,** including revision of the engagement rankings given to online content, monitoring and recommendation algorithms. This can only happen through public, auditable assessments of the risks of online services and the steps companies have taken to mitigate those risks. At minimum, this assessment would include the likelihood of the service to promote or accelerate discrimination and misleading content; and
 - ii. remove from the economic models of online services any incentive to promote or amplify discrimination, hate speech, or harmful misinformation. Neither the bad actors who create the disinformation content, nor the platforms should benefit from advertising revenue streams from ads served alongside disinformation. Consistent bad spreading disinformation should have their content down ranked from the engagement rankings of recommender and search algorithms (with full transparency of any action taken by the platform); and
 - iii. moderate online systems with algorithms that have adequate data sets and the predictive capacity to recognise hate speech and verified disinformation; and
 - iv. **disrupt disinformation networks** that mislead people and potentially change public opinion, by identifying and exposing coordinated online disinformation networks that deploy fake accounts and mislead people with content that aggravates existing fault lines in respective countries.
- 2. Unprecedented transparency online platforms and service providers should:
 - a) Ensure transparency of:
 - i) **moderation** decision-making by online platforms, informing users, account holders, and content creators about the reasons why their content was downgraded or removed; and
 - ii) **disinformation network campaigns** so that users learn about coordinated online disinformation networks that deploy fake accounts and mislead people ; and
 - iii) harmful content or platform manipulations users have been exposed to so that users learn about coordinated online disinformation networks that deploy fake accounts and mislead people; and
 - iv) the way in which platform algorithms recommend and moderate content online is provided to users, researchers, and regulators. See Annex I for Reporting and Transparency to combat disinformation during operation.

- b) Direct users to authoritative news sources and label at the point of misinformation – correcting the record in this way does NOT backfire by reinforcing misinformation, but empowers service users to access multiple reliable sources of information. See Annex I for more detail on this fundamental risk mitigation measure.
- c) **Support a sustainable fact-checker ecosystem**, with funding independent of contractual obligations, as fact-checkers can provide vital data sets by which the platforms can judge and redesign their algorithms.
- d) **Empower civil society organisations** (CSOs) to provide evidence to counter online platforms' failings and systemic risks by integrating them into each stage of the goals to foster technology that supports democratic rights. Civil society organisations must be given adequate access to platforms' data to allow public scrutiny.

3. Give back control of users' data to the people - A revolution on the culture of data collection is needed, so that platforms stop seeking consent to hoover up data on a user's every move and thought through a tick box to enter the service, but work to build their users' understanding of what categories of data are stored, what uses the data is put to, how tracking of behaviours and data aggregated from other services inform the algorithm's selections. Platforms must commit to

- a) Provide active user choice over what data they have to provide to a service and fully inform users about the kinds of data collected and inferred about them
- b) Give users the right to be fully informed about the kinds of data collection used to target them with the content they see.
- c) Provide active user choice over which algorithm serves them, and compel platforms to provide users with enough information to make a meaningful choice.
- d) Users should also have the right to reject algorithms based on user profiling at all, and have the option to turn off algorithmic selection, or to choose which elements of it they wish to enable.
- 4. Effective enforcement must include commitments from Governments and regulators to:
 - a) Ensure regulation is **robust and independent** and respects Articles 19(3) and 20 of the International Covenant on Civil and Political Rights (ICCPR).
 - b) Use of co-regulation in the form of Codes of Practice can be valuable, but only where a realistic co-regulatory backstop exists allowing incentivisation of commitments promised by platforms, and penalties to be applied in the event of failure to observe the relevant code.
 - c) **Resource and train regulators** to keep up with technological developments for example how artificial intelligence is used in platforms.

Section 3: Regulating AI

It is clear that AI can be of benefit to humans. However, AI has also been shown to pose various and multiple risks. I believe we should try to learn from some of the mistakes of social media and make preparations NOW to protect ALL human rights, before further damage is caused. –James, Avaaz member from Sweden

Artificial Intelligence is almost a misnomer, as current AI does not yet think for itself. It is as good or bad as the data that goes into it, and the uses to which it is put. As such, it is clear that AI, when deployed poorly, can damage democracies, accelerate <u>repression</u>, <u>violate</u> <u>privacy</u>, and exacerbate discrimination and <u>bias</u>.

These issues need urgent, integrated, and systemic policy solutions, and a narrow view of risk minimisation simply won't do. AI regulation needs to safeguard our human rights, protect democracy and care for our well-being. The race we have just witnessed to release generative large language models, like ChatGPT and Bard show how fast these systems can evolve past their intended use. These models, within weeks of release, were accused of creating texts that <u>defame</u> and <u>disinform</u>, with <u>later models having more propensity to create misinformation</u> than earlier ones.

Principles:

- 1) Human rights are the gold standard Regulating AI requires incentivizing safe design and operation that safeguards human rights, democracy, and well-being while involving stakeholders in risk assessment and mitigation. In cases where AI poses a proven risk to human rights, it should be banned.
- 2) **Transparency is key** this includes the mandatory disclosure of data collection and decision-making processes and an explanation of these processes that can be understood by those subject to them.
- 3) Informed and empowered human oversight is not a replacement for safe design but is needed to monitor all uses of AI that pose risks of harm to health and safety or fundamental rights.
- 4) Accountability for harms. Given the non-transparent nature of AI systems, accountability mechanisms should include adopting strict liability for harms committed through use of AI.
- 5) Enabling citizens to stand up to AI with explainability, simple processes to lodge complaints with a national authority and access to justice with a strict liability regime for harms caused by AI.

Key commitments and actions to achieve these principles include:

1) Human rights are the gold standard

a) Risk assessments and in particular Human Rights Impact Assessments (HRIAs) should be conducted by the developers and those deploying AI that they have acquired ("users of AI") before deployment to assess if the AI poses risks of harm to health and safety or a risk of adverse impact on fundamental rights.

One strong model for regulation of AI that has emerged in the past months is that of the HRIA. HRIAs come with a tested framework for assessing the risks that emerging AI systems pose. They are grounded in well-established international human rights standards and involve consultation with independent experts, civil society organisations, and subject groups. It is no wonder that the Dutch government² adopted HRIAs as a safeguard to help prevent serious AI harms, such as those that occurred in its child benefits scandal, from happening again.

HRIAs should be undertaken ex ante and form the basis of regular monitoring audits once the system is in use, as risks manifest during different stages of the AI lifecycle. Companies can investigate these potentially adverse consequences of their AI closely, adjust data sets and build mitigation measures, such as containment strategies, labelling, and compensation schemes into deployment plans.

See Annex II for examples of the type and nature of questions for an HRIA for AI use.

b) AI with proven risks to human rights should be banned,

Commitments should be given by regulators and developers to ending the use of AI in the following circumstances:

- i) Al practices that exploit vulnerabilities of specific groups based on characteristics such as age, physical or mental disability, gender, sexual orientation, ethnicity, race, origin, migrant, refugee, or asylum seeking status, should be prohibited.
- ii) The use of Al to infer emotions of a natural person, except for specific health or research purposes, should be prohibited due to the evidence of bias against minorities.
- iii) Banning AI profiling and risk-assessment in law enforcement and in the migration context: the impact of AI in exacerbating existing societal inequalities is only just beginning to be felt. Important evidence of these impacts is being produced by the scientific community, legal communities, and civil society organisations,

² https://www.humanrights.dk/tools/human-rights-impact-assessment-guidance-toolbox

demonstrating the impact of bias within automated decision making, surveillance, and other AI applications. Accordingly, as the use of AI in the migration context may lead to unequal treatment and exacerbate disparities among vulnerable communities, it should be banned for public and private actors. The potential for unconscious bias in algorithms used in this context should be addressed through regulation and review.

c) Al surveillance should be reduced and limited - Facial recognition software and emotion-recognition systems can be <u>biased against minorities</u>. Further, its widespread use, from street to online use, would be a pervasive breach of the human rights to privacy and to a private family life. Accordingly, real-time (live) and post (recorded) remote biometric identification systems should be subject to strict assessment and not permitted unless it can be demonstrated that its use carries no risks to health, security or human rights. Prohibitions should be put in place for its use in predictive law enforcement and immigration contexts.

Governments, regulators, and private companies must commit to a framework to:

- reduce and limit the use of remote biometric identification systems, such as facial recognition, in public spaces due to the high risk of intrusion into individuals' private lives. Such systems should be prohibited for use in
 - (1) law enforcement and
 - (2) commercial or administrative contexts.
 - (3) These prohibitions should extend to automated recognition in public spaces of human features, such as faces, fingerprints, DNA, voice, keystrokes, and other biometric or behavioural signals.

d) Supporting workers' rights

Commitments from employers and a regulatory system to enforce them are needed when AI management systems are used including:

- i) a legal duty for employers to consult trade unions on the use of high risk and intrusive forms of AI in the workplace;
- ii) workers should be aware of the AI systems at the workplace;
- iii) workers have the legal right to have a human review of decisions made by AI systems about them;
- iv) workers have the legal right to be able to "switch off" from work to have proper downtime in their lives; and
- v) an annual conformity assessment should be undertaken to address any bias arising in recruiting and management AI.
- vi) The regulatory system should be part of the national labour law and working conditions regulation with full recognition of so-called gig workers to full employment rights.

2) Transparency is key

e) Transparency through open auditing of AI against standards for AI that are based on human rights and pay equal attention to the technical and social aspects of the operation and impact of the AI.

The risks of AI to date have largely emerged as a result of the interplay between intended use, the context of deployment, and unintended effect of poor quality data sets.³ AI systems are deeply socio-technical, and a focus only on technical issues would fail to incorporate both problems and possibilities for system improvement.

An end-to-end, socio-technical approach generates process, mitigation strategies and documentation that improves system accountability, organisational memory, and compliance with proposed AI and data regulations. For those acquiring and incorporating AI systems into their operations, an audit provides crucial information that enables due diligence and proper assessment and comparison of the characteristics between different systems and vendors. The rigour should also underpin the standards of conformity bodies assessing AI and the ongoing monitoring obligations of high-risk AI users.

Al audits are useful for regulators and society, who can use audit reports to assess how systems work and their impacts. They are also useful for those developing and acquiring Al systems. We have attached a sample audit, relating to the right of non discrimination, developed by Eticas, in Annex III.

3) Informed and empowered human oversight

a) Commitments should be given by those using AI to implement human oversight across all AI systems that pose risks of harm to health and safety or a risk of adverse impact on fundamental rights. But we would like to draw attention to the fact that human oversight is not a panacea for fundamentally risky or harmful AI. The issues of over-reliance on technology and trust that it has the right answers, also called **automation bias**, has been frequently documented. So human oversight that is

³ See, e.g., Human Rights Impact Assessments for AI: Learning from Facebook's Failure in Myanmar (2021), <u>https://carrcenter.hks.harvard.edu/files/cchr/files/210318-facebook-failure-in-myanmar.pdf</u> (detailing how complex the interplay between intended use, context and data sets can be. The report states that, "Had Facebook conducted a baseline assessment for potential human rights impacts prior to operating in Myanmar, such indicators would have served as an early warning. The potential human rights impacts derived from the legal, political, and information landscape, including the government's discrimination and marginalisation of the Rohingya, would have been apparent from any human rights analysis of the market context prior to Facebook's entry.").

intended to address potential rights violations and harms to health or well-being must be performed by a workforce that:

- i) is properly trained;
- ii) has decent, reasonable working conditions including time to think and assess responses carefully; and above all,
- iii) has the authority and institutional backing to say no to decisions of AI, which potentially carry risks to rights of well-being.

4) Accountability for Harm

a) Ethical standards for AI

Regulation must involve a commitment to observe and cooperate in the production of internationally-recognised standards for the design and auditing of AI, including general purpose AI. These standards should for example cover the ways in which those systems collect and use the massive data sets on which the AI is trained is the first step to AI accountability, as well as all the moments of bias and other human rights impacts that arise through the operation of the AI in its specific context. For example, please see above and Annex III for proposals on an audit that could be used in developing a standard to identify and mitigate tendencies of an AI system to discriminate.

These standards must be independent and development must involve those subject to AI, and relevant civil society representatives. Observance of these standards should be incentivised by all parties to the compact, regulators, developers and funders. Cooperation with these regulatory standards must become standard behaviour for AI developers.

- **b) Strict Liability regimes -** Given the non-transparent nature of AI systems, accountability mechanisms should adopt strict liability for harms committed through use of AI.
- c) Compliance to the higher legal standards for importers and exporters of AI There should be no export of AI which would breach the developers own national AI rules, even though they may not breach the country of destination where the AI is to be used. Nor should there be import of AI that would breach AI development or use rules in the country of destination, irrespective of where it was developed.

5) Empowering citizens to stand up to AI

Empowering citizens is not just a matter of justice, but is also needed to create trust. Regulation needs to make sure that the AI systems brought to market are safe and trustworthy, but the public also expects the 'safety net' in the form of liability for cases where harm has occurred.

Often, people are not even aware an AI system has been involved in a decision made about them. This, combined with the complexity of AI systems, can make it almost impossible for citizens to stand up for their rights. To resolve this the compact should provide for:

- a) **Explainability** An individual must be able to ask for an explanation, which they can understand, of how AI made a decision about them, so that they can contest it, especially when the decision has serious impacts on their life.
- b) **Simple processes to lodge complaints** with a national authority, where rights have been violated by the use of an AI system,
- c) Allowing representation of those injured by relevant civil society bodies, if needed.
- d) Strict liability ie the presumption of fault on the part of the AI if harm has resulted from the use of the AI.

I'm more than worried about our future, especially when it comes to AI. In my work with young children, I see our future politicians, lawyers, nurses, and teachers. I wish and hope that their future will be as safe and sound as possible when it comes to protecting their privacy and that they will experience fairness and non-discrimination when it comes to laws concerning AI.

We are at the beginning of the era of AI and this is the right time to think carefully about the consequences before acting! Please do not miss this opportunity to ensure AI respects human rights!

Maria, Avaaz member from Finland

For further discussion of these additional comments please contact <u>Sarah.Andrew@Avaaz.org</u> or <u>christine@avaaz.org</u>

Annex I - Reporting and transparency to combat disinformation during operation

The output of the algorithm should be assessed against the following measures:

- The scale of disinformation on the platform including:
 - The number and frequency of user reported breaches of platform standards, specifically reports on hate speech and disinformation;
 - The number and frequency of disinformation detected through the platform's moderation algorithms and/or reported by reputable fact-checkers;
 - The reach of disinformation on the platform every platform can model the reach of a particular piece of content and this data should be provided to the auditors:
 - The numbers of fake accounts detected and removed;
 - The extent of unlabelled bots on the platform
 - Other patterns of inauthentic behaviour.
- The efficacy of the platform's algorithms to detect and mitigate breaches of the platform's standards. This would include for example:
 - The **speed** with which all reports are **assessed** by the platform's moderation algorithms and or human moderators;
 - The **speed** at which a **correction** is placed on misinformation content, and a comparison between the reach of the misinformation and the amount of views on the correction.
 - The **nature of any other action** taken in response to all reports, and the speed with which it was taken;
 - The **reach** of any correction the platform provided alongside the data on the reach of a particular piece of content
 - The **degree** to which the control measures in the algorithm **downgraded and suppressed promotion** of a given piece of disinformation;
 - The **removal of account**s spreading illegal material such as hate speech;
 - The **ability of the algorithm to detect repeat attempts** by such account holders to game the system by creating new accounts;
 - In the case of a user reported breach, the **communication of action taken to the user** who reported it; and
 - Where a breach could affect the rights and/or well-being of a wide set of users **on issues of public interest** for example content that stirs up hatred against immigrants, false claims about a crucial story in an election, or bogus claims on a cure for Covid-19, then **communication of the breach and the action taken** against the account holder's who created it, should be provided to **all affected users within the platform**, not just to the person who reported it. This is the only way the public can gain insight into the scale and

organisation of disinformation on the platform. Avaaz's full policy position on Correct the Record along with a study showing the need for corrections to be disseminated to all who were exposed to disinformation can be found here: Avaaz White Paper: Correcting the Record https://secure.avaaz.org/campaign/en/correct_the_record_study/

Setting key performance indicators (KPIs)

Finally, any audit will be more effective when cross platform KPIs are developed. Our understanding of which KPIs are most crucial will only develop across time, as the transparency afforded from the audit process described above will clarify which measures are most effective at scale. What's important is that the KPIs should be set by the independent auditor, with the public interest as its goal, not as at present, by the commercial platforms themselves. We suggest the KPIs should include:

- Number of instances of manipulative behaviour (fake accounts, impersonations, coordinated inauthentic behaviour, etc) on platform;
- The percentage of activity generated through such instances of manipulative behaviour on the platform as a proportion of all activity in a given country;
- Numbers of accounts or domains barred from participation to advertising or monetisation
- Ads removed or prohibited from their services
- The reach of disinformation on the platform every platform can model the reach of a particular piece of content and this data should be provided to the auditors, alongside confirmation of the reach of any correction the platform provided;
- Reach of corrections on platform;
- Reach of low-quality information on platform;
- Reach of hate speech on platform, items deleted; and
- Robustness effectiveness of steps taken to counter disinformation.

Annex II: Examples of questions that could be included in an AI HRIA questionnaire⁴

- How could the proposed deployment of this AI system potentially negatively discriminate against people on the basis of any of the following grounds (non-exhaustively): sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation?
 - a. Have you put in place processes to test and monitor for potential negative discrimination (bias) during the development, deployment and use phases of the AI system?
 - b. Have you put in place processes to address and rectify potential negative discrimination (bias) in the AI system?
- 2. Does the AI system respect the freedom of expression and information and/or freedom of assembly and association?
 - a. Have you put in place processes to test and monitor for potential infringement on freedom of expression and information, and/or freedom of assembly and association, during the development, deployment and use phases of the AI system?
 - b. Have you put in place processes to address and rectify potential infringement on freedom of expression and information, and/or freedom of assembly and association, in the AI system?
- 3. Are the relevant rights-holders and/or their legitimate representatives involved in the assessment of impact severity? Does the assessment of severity reflect the views of the relevant rights-holders?
- 4. Is there an operational-level grievance mechanism in place that contributes to ongoing impact management, as well as the identification of unanticipated impacts? If not, does the impact management plan include the establishment of such a mechanism?
- 5. To what extent has use of the AI system in a similar deployment already caused harm to health and safety or adverse impact on fundamental rights, or has given rise to significant concerns in relation to the materialisation of such harm or adverse impact, as demonstrated by reports or documented allegations submitted to national competent authorities?
- 6. To what extent are potentially harmed or adversely impacted persons dependent on the outcome produced by proposed AI use, in particular because for practical or legal reasons it is not reasonably possible to opt-out from that outcome?
- 7. To what extent are potentially harmed or adversely impacted persons in a vulnerable position in relation to the user of the AI system, in particular due to an imbalance of power, knowledge, economic or social circumstances, or age?

⁴ The Danish Institute for Human Rights' <u>Human rights impact assessment guidance and toolbox:</u> EU High-Level Expert Group on AI, Assessment List.

Annex III

An AI Audit proposal to identify Moments of Bias

Contents

Introduction	14
The purposes of auditing	14
Socio-technical algorithmic audits	15
Table 1 Summary Moments and Sources of Bias	15
The Process	16
a) The model card	16
Table 2 Model card fields	17
b) The System Map	18
c) Identification of Moments of bias	19
Table 3 Moments and sources of bias in pre-processing:	20
Table 4 Moments and sources of bias in-processing:	21
Table 5 Moments and sources of bias in post-processing:	22
d) The life cycle of mitigation	23
Table 6 The life cycle of mitigation measures	23
e) Mitigation measures for identified bias	24
Table 7 Sources of bias, testing for audit reporting and mitigation measures	25
f) The audit report	28
Notes and Acknowledgements	29

Introduction

The purposes of auditing

Algorithmic auditing is a way to inspect AI systems in their specific contexts. It is an approach and methodology that allows for a dynamic appraisal of regulation, standards and impacts. It is an essential tool for the transparency and accountability of AI.

This audit is based on the concept that AI is already an existing system. This audit methodology does not prompt a reflection on whether a system should exist in the first place. This should be considered through the use of a full Human Rights Impact Assessment and this audit anticipates that this is a key stage for all high risk uses post the introduction of the EU AI Act. This innovations in this annex were created by Eticas consulting, and are presented jointly with Avaaz Foundation.

Socio-technical algorithmic audits

As an end-to-end, socio-technical algorithmic audit inspects a system in context, looking at the specific data used and the constituencies impacted. It is end-to-end because it recognizes that algorithmic systems work with data produced by complex and imperfect individuals and societies, and operate and intervene in complex social and organisational contexts. An audit must also cover and provide consistency with legislation on human and data rights where they exist as well as labour rights, anti discrimination law. Where these rights do not exist in enacted legislation, their terms should be seen as standards to be matched or outreached.

Thus, Ai systems are deeply socio-technical, and a focus on technical issues would fail to incorporate both problems and possibilities for system improvement and impact testing that go beyond in-processing. The audit proposals here are focused on bias assessment, but not limited to it. See Table 1 for a survey of relevant possible bias such an audit might be designed to address.

The methodology to carry out such an audit must incorporate questions related to broader social impact and desirability, as well as the incorporation of end-users in the design process and the existence of recourse mechanisms for those impacted by algorithmic systems. For a system to pass an algorithmic audit, issues of impact, proportionality, participation and resource must be tackled.

Table 1	AI	Pre-processing	In-processing	Post-processing
Summary	life-cycle			
Moments and				
Sources of Bias				

Table 1 Summary Moments and Sources of Bias

Moments of bias	World → Data Data → Population Population → Sample Sample → Variables + Values	Variables + Values → Patterns Patterns → Predictions	Predictions → Decisions Decisions → World
Sources of bias	Techno-solutionist bias Selection bias Historical bias Sampling/generalization bias Population bias Survey bias Seasonal bias Survivor bias Rescue bias Exclusion bias Oversimplification, partial or biased featurization Omitted variable	Over and underfitting Measurement bias Hot hand fallacy Privacy bias Aggregation bias	Benchmark test bias Data visualization Automation bias Deployment bias

The Process

The process proposed for this audit in summary consists of

- a) The model card
- b) The system map
- c) Identification of moments and sources of bias
- d) The life cycle of mitigation
- e) Specific mitigation of identified bias
- f) Reporting

a) The model card

Model cards are documents designed to initially compile information about the training and testing of the AI, as well as the features and the motivations of a given dataset or algorithmic model.. The following fields should be considered:

Table 2 Model card fields

General information	Information on process	
 o Development phase Idea; Design; Development; Beta/Pilot; Production o System architecture High-level description of data architecture and dependencies, including process, system and model/s 	Information on processoDescription of purpose and role/service providedInformation on whether it is a new role/service or the automation of anexisting role/service. Basic overview of the purpose of the algorithmictool. It should include how and why the algorithmic tool is used. Purposeof the tool in terms of what it has been designed for and what it has notbeen designed for. This can include a list of potential purposes that thetool was not designed for: this can help to avoid misconceptions aboutthe scope and purpose of the tool. Description of benefits. Frequencyand scale of use.oStakeholder involvementDescription of any stakeholder consultation processes performed,including UX studiesoOrganizational contextHow the algorithmic tool is integrated into the decision-making processoHuman role/soRisks	
Information on training/validation data	Information on the model	
o Data collection methodology/origin Including time frame and geographical coverage of all data used, and APIs. Data collection process. This can include details on the original purpose of data collection and the context in which it was initially collected. If the data was initially collected for different purposes and is now being repurposed, information on legal basis. o Data types o Data quality certifications/procedures/roles o Datasets o Data formats	o Method/s used and justification Linear regression, logistic regression, decision tree. SVM algorithm, Naive Bayes algorithm, KNN algorithm, K-means, random forest algorithm o Simplified output/s Score, tag, categorization, recommendation, ranking o Variables Description of the types of variables or features used to train, test or run the model - for example 'age' or 'address'. In certain cases, it might not be feasible for a team to disclose all the variables in a dataset. In this case, teams should disclose - at a minimum: whether the data contains personal and special category information; variables of interest, such as protected characteristics and potential proxies; and variables with high predictive power or that have a significant bearing on the model. o Parameters	

	o Rules o Objective function/s and metrics Accuracy metrics (such as precision, recall, F1 scores), metrics related to computational efficiency.
Information on bias and impacts (in lab/operational settings)	Information on redress
 Error, FR and FN rates Profiling categories Vulnerable categories Selection rate per category and profile Average score for individuals and per profile in each category Impact ratio per category and profile Reverse-engineering results 	 Redress or review Cases performed, reviewed and review outcomes over time

b) The System Map

The system map then brings together information on the model, the system and the process. A first version can be designed by the auditor/s following the information provided in the MC, to be completed and validated by the development team/s. In this way

- i) *Model*: The model is the trained algorithm, that is, the rules adapted to a particular domain, which constitute the foundation of the technology we audit. Models are subject to performance evaluation, test, and can be compared to each other via benchmark datasets. The model is the core of an AI system, but it usually relies upon other elements (e.g. data pipelines, visualisation platforms,...) for it to work.
- ii) *System:* The system in this case refers to the entire technology. For a mobility service it could be the app that integrates a Machine Learning (ML) model to predict demand and adjust pricing, including the UI, including for example the data pipelines and protocols.
- iii) *Process*: By process we define the entire lifecycle of any unit of work, from the moment it enters into the workflow all the way to the decision and, if part of the process, the actual way the decision is utilised.

c) Identification of Moments of bias

Bias refers to a deviation from the standard. As such, and in technical terms, bias may be needed and desirable. In the context of AI accountability, however, "bias" has become an hypernym or umbrella term for lack of fairness and discrimination in data processes which result in individual and/or collective harms. By identifying and mitigating bias, we can ensure or protect fairness in AI systems.

Bias is the result of many factors, social and technical: from systematic errors introduced by algorithmic design choices, dirty data, sampling procedures, reporting protocols, or wrong assumptions that cause a mismatch between the input features and the target outputs. To date, *most studies on bias have focused on historical and aggregation bias*, that is, the need to identify protected groups and calculate disparate treatment and impact. This is at the heart of this methodology. However, bias and inefficiencies can emerge at other times, and a focus on historical and aggregation bias alone will lead to incomplete and therefore harmful assessments of bias. This will result in rights violations, stereotyping, bad or inefficient decisions, discrimination of individuals and groups, and the reproduction of processes of inequality and dispossession. Partial or wrongful identification of bias sources and inadequate mitigation measures will lead to unacceptable societal harms and compliance risks.

This audit distinguishes between moments and sources of bias. This provides the auditor with an overview of the possible causes of a given *disparate impact*, understood not only as an individual function of accuracy or performance but also as a general measure of (lack of) fairness in an *algorithmic process*. This audit method defines and identifies moments and sources of bias, establishes the documents and tests needed to assess compliance with legal and social requirements, provides an opportunity to address and mitigate inefficiencies and harms, and provides a measure for overall system fairness and impact.

Existing documentation, access to development teams and end users and the ability to test data processes will determine the auditor's visibility over existing an algorithmic system. Once implemented, the taxonomy of moments and sources of bias provides a systematic approach to identifying the root-cause of bias and harms, addressing them and documenting the process.

The categories proposed are not mutually exclusive nor relevant in all cases; however, identifying and characterising each one as distinct makes them easier to tackle and ensures a comprehensive assessment of all potential moments of bias. See Table overleaf

Moments of bias	Sources of bias	Definitions		
World → Data	Techno-solutio nist bias	Failure to consider no-tech or low-tech options, to perform a proportionality assessment or to take into account social and environmental issues before deciding to develop or implement an algorithmic system.		
	Selection bias	Due to problematic problem definition or data availability, the datasets selected are not related or targeted to the issue at hand, and this solve or address different problems.		
	Historical bias	Existing bias in the world that percolates into the dataset used for training, validation, and testing. Even if data is accurate and perfectly measured and sampled, the world as it is or was may lead to a model that produces harmful outcomes.		
Data → Population	Sampling/ generalization bias	Stems from non-random sampling of groups due to a limited or uneven sampling context (i.e. data about some groups is not available or certain sub-groups are not included). If all relevant members of data are not representative of the target population, it is possible that the trends identified for one population fail to generalize to new populations and don't accurately represent the environment the program is expected to run in.		
Population → Sample	Population bias	Arises from differences between the actual usage population and the design target population of a system. This means that the target population defined during the design and development phases is not representative of the population that will use the system after it is deployed. Population bias results in non-representative data and results that fit only the most salient groups while harming all minority groups.		
	Survey bias	Can occur when the data collection process relies on responses from interviews, surveys, and questionnaires, and the feedback received is either incomplete, inaccurate, or inconsistent due to social bias (preventive self-correction patterns to conform to the mean), recall bias or self-selection bias.		
	Seasonal bias	Emerges when failing to acknowledge the effects of seasonal patterns during the problem definition stage.		
	Survivor bias	A form of sample selection bias that only considers "surviving" observations, failing to consider those that ceased to exist. Relevant as many AI systems are used on vulnerable populations.		
	Rescue bias	A deliberate attempt to evade evidence that contradicts expectations. Consists of discounting data by selectively finding faults in a study or experiment, or by discounting faults when the data are viewed favorably.		

Table 3 Moments and sources of bias in pre-processing:

	Exclusion bias	The removal of certain valuable features or information during the post-randomization of the data.
Sample → Variables + Values	Oversimplificati on, partial or biased featurization Omitted	algorithm will miss relevant correlations, patterns and data. If some information is not adequately captured, unwarranted conclusions can be drawn from partial featurization. If some features are indirect proxies for other bias-prone features, bias will emerge. When one or more important variables are not included in the model, resulting in biased
	variable	regression coefficients and inaccurate statistical results.

Table 4 Moments and sources of bias in-processing:

Moments of bias	Sources of bias	Definitions
Variables + Values → Patterns	Over and underfitting	Overfitting (or high variance) occurs when a hypothesis function fits the data available perfectly but lacks predictive capacity for new data. In this sense, complex functions that capture many curves unrelated to the data or large neural networks with more parameters are more prone to overfitting. Underfitting (or high bias) occurs when a hypothesis function maps poorly onto the data available. Too simple functions or a model with not enough features is more prone to high bias.
	Measurement bias	Derives from a mismatch between training and target data types and tools (e.g. an image recognition system trained on images with a given resolution, when real-life inputs have lower resolution)
	Hot hand fallacy	Thinking that a model will continue to perform well because it performed well in the recent past without further justification or testing
	Privacy bias	Privacy-enhancing measures can impact both auditability and analysis. When implementing the data minimization principle, teams may chose to not collect attributes that the model may nevertheless infer, thus causing disparate impact that is difficult to identify and control. Additionally, differential privacy training and pruning may improve privacy but also reduce the influence of underrepresented data on the model, thus leading to worse model performance on that data
Patterns → Predictions	Aggregation bias	When a given model is not optimal for any group, or is skewed towards the dominant population. This type of bias is also known as ecological fallacy, for it occurs when incorrect or false conclusions are drawn about individuals by observing the population. Two types of aggregation bias need to be checked: Simpson's paradox (a statistical effect that occurs when trends are observed in different groups of data but such trends either disappear or reverse when groups are combined. An example of Simpson's paradox is the seeming equality in job opportunities for men and women at a larger scale that hides

unjustified inequalities for women when it comes to accessing certain jobs) and Modifiable
Areal Unit Problem (when geospatial unitary metrics are modeled at different levels of spatial aggregation. In such a case, different trends can be observed for the same values
depending on the aggregation scale).

Table 5 Moments and sources of bias in post-processing:

Moments of bias	Sources of bias	Definitions
Predictions → Decisions		
	Data visualization	Visualization can introduce bias into the algorithm through the Framing Effect, a type of cognitive bias where people's choices depend on the ways in which options are presented to them and may be triggered by data visualization choices. Visualization choices may also be related to dark patterns (a user interface designed to trick users into doing things thay may not be in their best interest). Other sources of visualization bias are Availability Bias (when the data used is the most readily available, although it may not be the most representative or the actual evidence for a given problem), Anchoring Bias (a cognitive bias that causes people to rely excessively on the first piece of information received) and Signal Error (when caveats or gaps in the data are overlooked to avoid complexity).
Decisions → World	Automation bias	The propensity for humans to favor suggestions from automated decision-making systems and to ignore contradictory information made without automation, even if it is correct.
	Deployment bias	Captures all instances where organizational, budget, technical or training issues may impact on how decisions are ultimately made and provided to those affected.

d) The life cycle of mitigation

An effective audit must also inspect and propose mitigation recommendations to algorithmic systems in an iterative process of interaction between the auditor/s and the development team/s. The method should provide templates and instructions to guide such interaction, specifying the data inputs that are necessary for auditors to complete the assessment and validate results. Crucially it should propose and record mitigation across each of the stages an AI system may impact on citizens, from pre deployment testing by the developer to post deployment by the user.

It works with systems that make decisions on individuals or groups based on known data sources, regardless of whether they use machine learning or classic computing. This definition includes most systems used by the public and private sectors to make decisions on resource allocation, categorization and identification/verification in sectors such as health, education, security, finance and for applications like fraud detection, hiring, operations management, or prediction/risk assessment.

Table b The life	AI life-cycle	Pre-processing	In-processing	Post-processing
cycle of mitigation measures		System discarded, re-designed or adjusted Randomization Re-training Re-labelling Remove sensitive attributes Synthetic data Rules Non-probabilistic purposeful sampling design Stratified random sampling Data augmentation	Regularization/de-regularizatio n Increase/decrease dataset size Add/remove polynomial features In-processing constraint optimization Homogenize testing methods Increase measurement frequency Test in lab and real Re-assess ground truth Document impact and trade-offs Safe collection of sensitive attributes	Testing with concept activation vectors Cross-validation UX review A/B testing Causal testing Staff training Clear roles and responsibilities for human in the loop Adversarial testing

Table 6 The life cycle of mitigation measures

Resampling Re-weighing PerturbationRe-calibration Re-testing Thresholding Explainability analysis and testingCausal testing Re-collectionExplainability analysis and testingData linkage Probabilistic matching Individual reference identifiersFairness metrics Confusion matrixHistorical analysis Complete data Confidence Internal and external review Thresholds for data deviations Add/remove BlindingRe-calibration Re-testing Explainability analysis and testing Explainability analysis Confusion matrix	d
Adversarial training and coding	

e) Mitigation measures for identified bias

Based on the documentation provided and access to team developers and the data available, different types of tests can be designed to determine whether different types of bias are impacting systems in ways that may cause harm to individuals, groups, society or the efficient functioning of an AI system. In all cases, bias testing involves a documentation and literature review, interviews with developers/implementors and a good understanding on who is impacted by AI systems and how. Bias testing involves statistical analysis and checking, and auditors have a choice of fairness definitions and metrics to choose from. Statistical notions of fairness such as those described by Verma & Rubin (2018) are a good starting point and can be the basis for more advanced approaches such as similarity-based measures and causal reasoning. In some cases, bias testing requires reaching out to end users or those impacted by systems.

The following table provides a non-exhaustive list of the documents that can be required/developed/assessed and the types of tests auditors can perform to gain knowledge of how a system works and its context. The right column lists the possible mitigation measures that can be implemented by system developers and auditors to control for negative externalities and

prevent instances of harm and unfair decision-making. The specific design of the relevant documentation and mitigation measures will need to be developed in the context of a particular use case, and it is not the responsibility of the auditor to implement mitigation measures. Hence, columns B and C below are fundamentally different: column B describes the methods an auditor may use to identify bias, while Column C points to the mitigation measures that the auditor may suggest are implemented to improve a given system.

Table 7 Sources of bias, testing for audit reporting and mitigation measures

Sources of bias	Documentation and testing for audit reporting	Mitigation measures
Techno-solutio nist bias	Documentation: data protection impact assessment if any; HRIA, proportionality test, CBA, problem definition, benchmarking, success indicators Testing: analysis of alternatives, assessment of developer expectations vs. performance in other relevant cases, assessment of desirability and compliance, assessment of environmental impact, assessment of objective function	System discarded, re-developed or adjusted
Selection bias	Documentation: HRIA, problem definition, target population, training data, labels and features, training dataset list and justification Testing: check all datasets for relevance, check labels and features against abstract constructs, identify and assess "inherited" datasets, assess data minimization compliance	Conduct a comparative assessment between the population represented in the training data and the target population Randomize sub-group selection Re-training with relevant, minimized data Re-labelling to ensure relevance
Historical bias	Testing: identification of protected groups based on the literature and analysis of the problem, training data and end users, validate Word embeddings (in NLP)	Recognize the existence of patterns of discrimination in the application and generation of data over time Remove sensitive attributes Use synthetic data Add rules to the model to ensure fairness
Sampling/gene ralization bias	Documentation: sampling methodology and justification, development sample	Use non-probabilistic purposeful sampling designs like quota sampling Use stratified random sampling and strategic sampling

	Testing: test performance of classifiers on under-sampled and control groups, assess instances of oversimplification	Data augmentation and resampling to achieve demographic parity Re-labeling to modify the proportion of positive instances across protected groups Re-weighing to indicate a higher frequency for an instance, to modify the importance of the instances, and to improve the stability of a classifier Perturbation to modify certain aspects of the data according to a given notion of fairness
Population bias	Documentation: cultural/language/geographical adaptation plans Testing: identify hidden populations, variability testing, test in different cultural/language/geographical contexts	Re-training to ensure debiased and comprehensive training data Causal testing
Survey bias	Documentation: data sources Testing: assess quality, completeness and robustness of data used	Re-design data collection Re-collection and re-labeling to account for survey bias Data linkage to contrast information about a given element from different sources Use probabilistic matching and individual reference identifiers
Seasonal bias	Documentation: problem definition Testing: assess relevance and impact of seasonal bias in data sources and processes	Historical analysis to identify seasonal patterns and adjust predictions Re-labeling to minimize seasonal variance
Survivor bias	Testing: identify hidden populations	Address incomplete data Use confidence intervals in the analysis Re-training to ensure debiased and comprehensive training data
Rescue bias	Testing: assess alternative explanations, review findings with end-users	Create internal review and validation process for all sampling decisions
Exclusion bias	Documentation: justification for discarded features, feature importance assessment Testing: check relationship between features and labels	Create internal review and validation process for all sampling decisions Establish thresholds for data deviations
Oversimplificatio n, partial or biased featurization	Documentation: problem definition, data discovery plan, features, list and justification for all variables Testing: review literature, review variables, check for missing relevant variables, check all variables for indirect bias	Create external review and validation process for all variables and features Add features to match complexity of context and problem Remove bias-prone variable or variables that may act as proxies for protected attributes Blinding Adversarial coding

Omitted variable	Documentation: pre-processing data plan Testing: assess discovery and pre-processing plan, use control variables	Re-training to ensure debiased and comprehensive training data Create internal review and validation process for all variables, features and labels Adversarial training
Over and underfitting	Documentation: data preparation process, data annotation guidelines, piloting results Testing: assess ground truth labels	Regularization/de-regularization Increase/decrease the size of the training dataset Add/remove polynomial features In-processing constraint optimization
Measurement bias	Documentation: description of real-life conditions and tools Testing: check for disparities between lab/real life tools and data (including lighting conditions in the case of computer vision/FR)	Homogenize testing methods Increase measurement frequency Test in lab and real settings before launch
Hot hand fallacy	Documentation: history of model versions Testing: check for recent event/dynamics that may impact model performance and impacts	Re-assess ground truth, variables and measurements before launch of major system updates
Privacy bias	Documentation: data minimization report and measures, differential privacy, pruning Testing: calculate impact on auditability and performance on protected groups	Document impact and trade-offs of privacy-driven decisions. Create mechanisms for sensitive attributes to be collected for auditing purposes but not data training. The auditor may act as recipient and custodian (processor) of such data.
Aggregation bias	Documentation: test and validation data, objective function, model performance metrics, benchmark datasets, evaluation metrics Testing: test model types, hyperparameters and optimization methods, check for differential measurement/accuracy across groups, calculate results/rates for each protected group identified in pre-processing, calculate and document impact ratio per group (results for protected vs results for most salient group) and individuals (results for individuals of a protected group vs. results for individuals in the most salient group), define/validate fairness definitions and metrics	Re-calibration and testing for all groups Regular testing for known discriminatory variables (gender, race, age, location, etc.)Thresholding Explainability analysis and testing Calculate fairness metrics: TP, FP, FN, TN, PPV, FDR, FOR, NPV, TPR, FPR, FNR, TNR DevelopConfusion $\boxed{\begin{array}{c} Actual - Positive \\ Predicted - \\ Positive \\ Negative \\ Negative \\ \end{array}}$ $\boxed{\begin{array}{c} Actual - Positive \\ Actual - Negative \\ True Positive (TP) \\ FDR = \frac{FP}{TP+FP} \\ FPR = \frac{FP}{TP+TN} \\ FPR = \frac{FP}{TP+TN} \\ FNR = \frac{FN}{TP+FN} \\ FNR = \frac{FN}{TP+FN} \\ TNR = \frac{TN}{TN+FP} \\ TNR = \frac{TN}{TN+FP} \\ TNR = \frac{TN}{TN+FP} \\ \end{array}}$ matrix:

Benchmark tests	Testing: test benchmark data for representativity across all relevant groups	Move away from accuracy as reference benchmark Testing with concept activation vectors (CAVs) Cross-validation
Data visualization	Documentation: all visualizations and interfaces Testing: framing effect, dark patterns, availability bias, anchoring bias, signal error	Create internal UX review process and validation for all visualizations and interfaces A/B testing (with humans in the loop and/or impacted communities) Use causal methods (such as directed acyclic graphs)
Automation bias	Documentation: role and training of "humans in the loop" Testing: staff interviews, measurement of instances of substantive human intervention	Staff training Clear attribution of roles and responsibilities (%) for systems and humans Create procedure for all human intervention to be justified and documented, ensuring accepting and rejecting algorithmic decisions creates similar burden on humans
Deployment bias	Documentation: process map, access to user interfaces, transparency policy, explainability documents, redress procedures Testing: staff and user interviews, adversarial auditing/tests	Develop deployment plan that includes budget breakdown, training descriptions, roles and responsibilities in algorithmic decision implementation Adversarial testing

f) The audit report

A crucial part of auditing is documentation, and so all interactions and documents exchanged must be compiled and either kept on file by system owners (and, if both parties agree, by auditors). **Audits should always result in a public document**. The basic structure of an audit report should include:

I. Background

II. Audit objective, scope and methodology

III. Audit results

IV. Audit follow-up

V. Acknowledgements

ANNEX I Status of audit recommendations

APPENDIX I Management response

An AI operator should anticipate producing three main audit reports:

a) Internal report with mitigation measures and annexes

This document captures the process followed, the issues identified and the mitigation measures that have been applied or can be applied. Contrary to financial auditors, algorithmic auditors do engage in proposing solutions, monitoring their implementation and reporting on the final results. The internal audit report need not be published.

b) Public report

Final version of the audit process, where auditors describe the system, the auditing methodology, the mitigation and improvement measures implemented and further recommendations, if any. The public audit report must also include a proposal for the periodicity and methodology/metrics to be used in follow-up audits.

c) Periodic reports

Follow-up audit reports. These must always refer and provide access to the initial audit report, if it is still relevant, and provide guarantees that the system developers have continued to test for bias, implement mitigation measures and control for impact. Depending on the complexity of the system/s, both parties may agree to produce an internal and a public version of each periodic audit.

Notes and Acknowledgements

On limitations:

- This methodology is a prototype.
- This methodology was primarily designed to audit supervised ML recommender systems. With minor amendments, it has been tested auditing ML supervised and semi-supervised recommender systems, computer vision/facial recognition systems and systems that use Natural Language Processing to make decisions about individuals.
- In some cases, the terms and definitions used in the methodology deviate from those used in some fields of knowledge. Socio-technical work requires flexibility as relevant fields have developed in isolation. To address this and avoid misunderstandings, definitions are provided in all cases to ensure auditors understand what each term refers to and how it is used in the context of algorithmic auditing.

Acknowledgements

This work is the result of 5 years of auditing algorithms with the team at Eticas Tech. During this time, numerous relevant contributions have been taken into account to ensure that the practices described draw from the experience of many other scholars, practitioners and relevant communities.