Written evidence submitted by the Ada Lovelace Institute (OSB0101)

We write to you from the Ada Lovelace Institute, an independent research institute and deliberative body with a mission to ensure data and AI work for people and society, regarding your examination of the draft Online Safety Bill and its goal of ensuring online platforms are safe for users. In light of the Government's plans to grant Ofcom new regulatory powers to meet this objective, we wish to highlight some of the important challenges we think this Bill must address and what specific powers Ofcom will need to meet these challenges.

In particular, the focus of this letter will be on the specific kinds of powers the Bill should articulate to create both meaningful powers of regulatory inspection of online platforms and a healthy ecosystem of auditing and assessment of these platforms.

We understand a core objective of the Online Safety Bill is to ensure that algorithms are not a driving force of harm. To achieve this objective, the Bill will need to provide **new powers** for Ofcom to **audit and assess** whether online platforms are meeting such standards.

This is the focus of our letter – **if the Online Safety Bill seeks to ensure that algorithms are not a driving force for harm, the Bill must provide more stringent and clear powers for Ofcom to carry out a regulatory inspection of platforms and the algorithmic systems used to deliver and moderate online content.** These powers should include:

1. Clearer powers for Ofcom to undertake regulatory inspections of online platforms when they deem appropriate.
2. Powers that enable Ofcom to perform technical audits, assessments and monitoring of platform behaviour, including algorithmic behaviour, whenever Ofcom deems appropriate.
3. Powers that create a healthy 'ecosystem of inspection' by enabling a marketplace of trusted independent auditors; empowering independent auditing and assessment from academic labs and civil society organisations; and granting Ofcom the power to penalise platforms that actively seek to disrupt independent auditing and assessment methods or refuse to conduct such audits.

The Online Safety Bill presents a unique moment for the United Kingdom to pioneer novel methods to ensure that algorithmic systems and online platforms are held accountable. While legislation that seeks to address online harms has already been rolled out in countries like Germany, and draft bills have been introduced in Canada, Australia and the European Union, there is currently no existing standard or clear practice for how regulators can meaningfully identify harms as they occur, or monitor efforts to address them.

In a joint paper by Reset and the Ada Lovelace Institute titled *Inspecting algorithms in social media platforms*,[1] a panel of experts on social media content moderation issues highlighted that current models of addressing online safety issues rely on insufficient forms of self-

[1] Ada Lovelace Institute and Rest (2020). *Inspecting Algorithms in Social Media Platforms*. Available at: https://www.adalovelaceinstitute.org/report/inspecting-algorithms-in-social-media-platforms/

regulation, and that these create information asymmetries between social media platforms, regulators and the public. In short, current models of addressing online harms rely on platforms to be good and honest partners. Despite their promises, platforms have routinely refused to share information on their policies, design decisions, and measures of effectiveness of automated and human moderation systems. While some platforms have relied on transparency reports that outline what kinds of content have been removed, these have come under criticism on the grounds that they self-select certain information and are not capable of being externally validated.[2]

Currently, regulators, civil society organisations, journalists and academic labs do not have the powers nor the access to the evidence to independently assess platform moderation issues. There is evidence that platforms act to exacerbate this deficiency by routinely shut down independent auditing and assessment efforts. Recent examples include Facebook shutting down independent auditing methods by AlgorithmWatch[3] and the New York University.[4]

These examples **demonstrate the urgent need for regulation that creates a healthier ecosystem of accountability for online platforms**. In this ecosystem, regulators would be empowered to check platform claims of online harms, and to audit and assess the prevalence and existence of potential harms as they arise. Likewise, regulation would empower civil society organisations and academic labs to perform independent audits and assessments of platform behaviour by enabling greater access to platform data.

Enabling this ecosystem of accountability for platforms would entail making amendments to the Online Safety Bill:

1. The Bill should provide clearer powers for Ofcom to undertake regulatory inspections of online platforms when they deem appropriate.

In our paper *Examining the Black Box*,[5] the Ada Lovelace Institute defined regulatory inspection as a broad approach for assessing an algorithmic system's compliance with regulations and norms of governance. Our paper *Inspecting algorithms in social media platforms* highlights that inspections must include regulators having the capacity to access three kinds of evidence:

1) Policies – company policies and documentation that relate to the kinds of harms they are moderating for (e.g. their policy defining hate speech that guides internal moderation teams assessment practices)

---

[2] Lyons, K. (2021). 'Facebook releases shelved content transparency report after criticism it wasn't being transparent'. *The Verge.* Available at: https://www.theverge.com/2021/8/22/22636508/facebook-releases-shelved-content-transparency-report-content-coronavirus

[3] Kayser-Bril, N. (2021). 'AlgorithmWatch forced to shut down Instagram monitoring project after threats from Facebook'. *AlgorithmWatch.* Available at: https://algorithmwatch.org/en/instagram-research-shut-down-by-facebook/

[4] Bobrowsky, M. (2021). 'Facebook Disables Access for NYU Research Into Political-Ad Targeting' *The Wall Street Journal.* Available at: https://www.wsj.com/articles/facebook-cuts-off-access-for-nyu-research-into-political-ad-targeting-11628052204

[5] Ada Lovelace Institute (2020). *Examining the Black Box: Tools for assessing algorithmic systems*. Available at: https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

2) Processes – assessment of a company's process for identifying and removing that content (which may involve interviews with staff members)
3) Outcomes – the ability to assess the outcomes of those policies, including the behaviour of algorithmic systems that amplify or moderate content.

Our belief is that Ofcom will need to have clearer powers to perform these kinds of inspections, for any platform that falls under their purview, and at any time of their choosing. Without these powers, Ofcom will be unable to assess the prevalence of harmful material on a platform.

2. The Bill should provide Ofcom with the ability to perform technical audits, assessments and monitoring of platform behaviour, including algorithmic behaviour, whenever Ofcom deems appropriate.

A regulatory inspection process may involve technical audits of a platform's behaviour. In a forthcoming paper, the Ada Lovelace Institute has surveyed the existing literature around methods for performing technical audits to identify and track online harms on major platforms. It is worth noting that our research discovered that the overwhelming majority of these audits were conducted by independent civil society organisations and academic labs, and none were conducted by industry labs themselves.

Our forthcoming research on technical methods for regulatory inspection of algorithmic systems **identified six common methods for technical audits of online platforms**, each of which can help regulators answer different kinds of questions. We detail these methods below, but note that some of these methods involve active monitoring of a platform's behaviour, a practice that the current Bill language does not enable.

Code audit	Auditors have direct access to the codebase of the underlying the system, or pseudocode (plain English descriptions of what the code does).	These audits are good for understanding the intentions of algorithms; in the case of machine learning systems, they're useful for understanding how objectives are being optimised. Regulators could, for example, ask platforms to provide pseudocode descriptions of a content recommendation system to inform whether their design is aligned with legislative codes of practice.
User survey	Auditors conduct a survey and/or perform user interviews to gather descriptive data of user experience on the platform.	These audits are useful for gathering information about user experience on a platform – to paint a rough picture of the kinds of problematic behavior that could then be further investigated. Regulators could, for example, poll user populations of interest, such as children. Ofcom used this method in a 2020 survey on COVID-19 misinformation sources.[6]
Scraping audit	Auditors collect data directly from a platform, typically by writing code to automatically click or scroll through a webpage and 'scrape' (or collect) data	These audits are useful for understanding content as presented on the platform and making descriptive statements (e.g. 'this % of search results contained this term'), or comparing results for different groups or

[6] Ofcom (2021). 'Covid-19 news and information: consumption and attitudes'. *Ofcom.* Available at: https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/news-media/coronavirus-news-consumption-attitudes-behaviour

	of interest (for instance, text that users post).	terms. Regulators could use this method either for one-off investigations into publicly available content, or to create and maintain a dataset over time that could be used for a range of inspection activities.
API audit	Auditors access data through a programmatic interface provided by the platform that allows them to write computer programs to send and receive information to/from a platform.	For example, an API might allow a user to send a search term and get back the number of posts matching that search term. This method provides easier programmatic access to data than a scraping audit, and allows easier automation of collection for descriptive statements or comparative work. Regulators could use information gathered via this method to assess the prevalence of content on a platform, or to set out standards for what kind of access the API should provide a regulator as part of their inspection.
Sock-puppet audit	Auditors use computer programs to impersonate users on the platform (these programs are called 'sock puppets'). The data generated by the platform in response to the programmed users is recorded and analysed.	This method is useful for understanding what a particular profile or set of profiles of users may experience on a platform. Regulators could use sock puppets to impersonate users from different demographics (for instance, under-18 users) to use the platform and record content recommended to them, to monitor the prevalence of harmful content.
Crowdsourced audit	A crowdsourced audit (sometimes known as 'mystery shopper') employs real users to collect information from the platform while using it – either by manually reporting experience or through automated means like a browser extension.	This method is useful for observing what content users are experiencing on a platform and whether different profiles of users are experiencing different content. Regulators could commission a panel of users in their jurisdiction to use a browser extension/custom browser to collect user data and analyse the content they see.

3. The Bill should create an 'ecosystem of inspection' by enabling a marketplace of trusted independent auditors; empowering independent auditing and assessment from academic labs and civil society organisations; and granting Ofcom the power to penalise platforms that actively seek to disrupt independent auditing and assessment methods or refuse to conduct such audits.

The current landscape of algorithmic auditing is not functioning well. Audits of algorithmic behaviour are currently a voluntary exercise that platforms may choose to complete, and in some cases have been shown to have been carried out in bad faith efforts to intentionally mislead regulators and members of the public.[7] Providing regulators with the power to conduct audits themselves would help address this issue, but regulators may also not have the capacity and resources to audit every firm at every given time.

To address this challenge, we strongly recommend the Online Safety Bill include elements

[7] Engler, A. C. (2021).' Independent auditors are struggling to hold AI companies accountable'. *Fast Company.* Available at: https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue

that enable a healthy ecosystem of audit and assessment. This would involve three ecosystem-focused changes:

a) **Enabling a marketplace of independent auditors that platforms can turn to.** There are already some for-profit auditing firms that specialise in delivering algorithmic audits of an algorithmic system's behaviour.[8] The Bill could help foster a new marketplace of independent auditors by granting Ofcom the power to mandate a platform to undertake an independent audit from a third-party agency that Ofcom has vetted and approved.

b) **Empowering independent auditing and assessment from academic labs and civil society organisations.** The Bill must recognise that regulators sit within, and will rely on, a wider ecosystem of inspection in which civil society organisations and academics are empowered to provide independent audits and assessments of platform behaviour. Many of the auditing methods we describe above break down because platforms do not provide relevant information or access to the data an auditor needs to perform these assessments. The Online Safety Bill could address this by granting Ofcom the power to compel platforms to provide certain data or APIs to third-party auditors who can undertake their own independent audits.

c) **Granting Ofcom the power to penalise platforms that actively seek to disrupt independent auditing and assessment methods or refuse to conduct such audits.** In discussions with independent auditors from civil society organisations and academic labs, many described their relationship with social media firms as one in which platforms treat them as adversaries rather than partners. In many cases, online platforms like Facebook have actively disrupted efforts to run these audits. To mitigate this threat to accountability, the Online Safety Bill should empower Ofcom with the ability to penalise or fine platforms that take steps to actively disrupt independent auditing capabilities.

Finally, we appreciate that a core challenge this Bill faces is defining a clear goal and objective for regulating online harms, and doing so early in a wave of pioneering legislation in a global context. Ensuring that algorithms are not a driver of harm is a laudable goal, but it will be no easy task to create specific definitions of these harms, or to determine what 'good' practices for online safety look like. Developing standards of appropriate platform moderation behaviour will take time, as will ensuring definitions of harm do not inadvertently create undesired chilling effects or risks to rights of free expression. While the Ada Lovelace Institute is not offering an answer to this extremely difficult question, our work to engage members of the public[9] in conversations around AI and data highlight the importance of using participatory methods of engagement to involve people in making determinations about how 'harms' are defined. We would strongly recommend the Bill includes a requirement that definitions of certain harms are made in part through the use of citizen juries or review boards that contain members of the UK public.

[8] Notable examples include ORCAA and Arthur.ai.

[9] See Ada Lovelace Institute (2020). 'The Citizen's Biometrics Council'. Available at: https://www.adalovelaceinstitute.org/project/citizens-biometrics-council/ and Ada Lovelace Institute (2021). *Participatory data stewardship.* Available at: https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/

Thank you very much for your kind consideration of this letter, we hope our recommendations around auditing and assessment powers for Ofcom prove useful and relevant to your examination of the Bill and we would be happy to provide more detail on any of these issues if helpful to the committee.

Yours sincerely,

Carly Kind, Director, Ada Lovelace Institute

About the Ada Lovelace Institute

The Ada Lovelace Institute is an independent research institute and deliberative body founded and funded by the Nuffield Foundation in 2018, with support from the British Academy, the Royal Society, the Royal Statistical Society and the Alan Turing Institute. Our mission is to ensure data and artificial intelligence work for people and society.

We do this by working with policymakers, civil society organisations, academic labs, industry organisations and members of the public to build evidence and foster constructive debate about how data and AI affect people and society. Our work focuses on five programme areas: 1) health and COVID technologies, 2) governance of biometrics data and technologies, 3) ethics and accountability in practice, 4) the future of digital regulation, and 5) public-sector uses of AI systems and data.

September 2021