Ada Lovelace Institute

Reset.

Inspecting social media algorithms
For Government, policymakers and researchers

Inspecting algorithms in social media platforms

As algorithms are designed and deployed at unprecedented scale and speed, there is a pressing need for regulators to keep pace.

On 6 August 2020, the Ada Lovelace Institute and Reset convened a group of international experts to identify the technical and policy requirements for inspecting algorithms in social media platforms, using the case study of COVID-19 misinformation. The workshop builds on previous work by the Ada Lovelace Institute on methodologies to inspect algorithmic systems, and by Reset on digital information market governance and the spread of information.

It was the first of three workshops convened by the Ada Lovelace Institute with partners in different domains, to not only develop thinking in each area, but also to identify shared needs, methodologies, challenges and solutions for the regulatory inspection of algorithmic systems across sectors.

This briefing presents insights from the workshop, and our corresponding recommendations for policymakers, with a focus on the UK and European Union.

Insights:

1. **The current model of self-regulation is insufficient, and cements information asymmetry between social media platforms and the public**

Currently, technology companies can launch, publicise and even reverse misinformation interventions at their discretion. External efforts document troubling gaps between companies' publicised interventions and the realities of COVID-19 misinformation on their platforms, but public authorities and other relevant third parties cannot access the evidence needed to analyse harms related to the platform.

2. **An algorithm inspection will require detailed evidence on companies' policies, processes and outcomes, and new methods of access to evidence**

Workshop participants identified the types of evidence – on policy, process and outcomes – they would need to analyse harms occurring on the platform, and the platform's expected behaviour in response to harms, and to verify platform claims about the role of algorithms in mitigating or increasing harms. They also suggested methods to access this evidence, from interviews with company staff to an inspector-specific API, many of which required some participation from technology companies.

3. Algorithm inspection brings with it significant opportunity but will require careful design to deliver on its potential

Governments must develop and enact a public policy agenda that regulates the digital marketplace, and aligns its interests with those of democratic and social integrity. At the same time, audit regimes must be proportional to the types of companies under review, and governments should anticipate and mitigate associated risks, including the potential for abuse.

Recommendations:

The regulator responsible will need:

1. Compulsory audit and inspection powers

An independent regulator should be empowered and resourced to enforce its obligations. This governance framework can only work on one condition: it requires transparency between the platforms and an independent regulator. The regulator should have the power to demand the granular evidence necessary to fulfil its supervisory tasks, and have enforcement powers when platforms do not provide that information in a timely manner.

2. Information-gathering powers that extend to evidence on policy, process and outcomes

The regulator must have the authority to request evidence on a social media platform's policy, process and outcomes, and technology companies will need to ensure they have capacity to respond to these requests, which could include methods such as interviews, APIs, or disclosure of internal policy documentation.

3. Powers to access and engage third-party expertise

An algorithm inspection requires a multidisciplinary skill set, although relevant expertise for any given inspection will vary based on context and industry. While the regulator should have some skills in-house, it will need the ability to access and instruct third-party expertise. This could include access by academics to conduct research in the public interest.

Introduction

As algorithms are designed and deployed at unprecedented scale and speed, there is a pressing need for regulators to keep pace with technological development; they must establish the systems, powers and capabilities to scrutinise algorithms and their impact.

There is no existing methodology for a regulatory algorithm inspection, although it is likely that any inspection will be guided by context, and its scope and function will depend on the industry and application under consideration. One such context is social media, where civil society organisations,[1] governmental bodies[2] and parliamentarians have already begun to call for algorithm inspections powers.[3]

A regulatory inspection of algorithmic systems, often referred to as an audit of algorithmic systems, is a broad assessment approach of an algorithmic system's compliance with regulation.[4] In the case of social media platforms and misinformation, this activity is forward-looking; the regulation in question is not yet in place.

In the UK, the Online Harms White Paper proposes a system of accountability for technology companies, including an independent regulator to oversee companies' compliance with a new set of rules.[5] It is also under discussion through the Digital Services Act in the EU. While the regulator's legal powers are still under discussion, the remit and capacity for algorithm inspection will be essential to effective oversight.[6]

On 6 August 2020, the Ada Lovelace Institute and Reset convened a group of international experts to identify the technical and policy requirements for inspecting algorithms in social media platforms, using the case study of COVID-19 misinformation. Participants included regulators, policymakers, and academic researchers from the social sciences, computer science and AI ethics.

[1] Demos, Doteveryone, Global Partners Digital, Institute for Strategic Dialogue, Open Rights Group. (2020) 'Algorithm inspection and regulatory access'. *Demos*. Available at: https://demos.co.uk/wp-content/uploads/2020/04/Algo-inspection-briefing.pdf (Accessed: 30 October 2020).

[2] Center for Data Ethics and Innovation. (2020) 'Review of online targeting: final report and recommendations'. *Gov.uk*. Available at: https://www.gov.uk/government/publications/cdei-review-of-online-targeting/online-targeting-final-report-and-recommendations (Accessed: 30 October 2020).

[3] Clarke, N. (2018) 'App developers should have products inspected by the state to make sure they're not racist, Labour'. *The Sun*. Available at: https://www.thesun.co.uk/news/7350592/app-developers-should-have-products-inspected-by-the-state-to-make-sure-theyre-not-racist-labour-say/ (Accessed: 30 October 2020).

[4] Ada Lovelace Institute and DataKind UK. (2020). 'Examining the black box: tools for assessing algorithmic systems'. *Ada Lovelace Institute*. Available at: https://www.adalovelaceinstitute.org/wp-content/uploads/2020/04/Ada-Lovelace-Institute-DataKind-UK-Examining-the-Black-Box-Report-2020.pdf (Accessed: 30 October 2020).

[5] UK Department for Digital, Culture, Media & Sport and UK Home Office. (2020) 'Online harms white paper'. Available at: https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper (Accessed: 30 October 2020).

[6] Demos, Doteveryone, Global Partners Digital, Institute for Strategic Dialogue, Open Rights Group. (2020) 'Algorithm inspection and regulatory access'. *Demos*. Available at: https://demos.co.uk/wp-content/uploads/2020/04/Algo-inspection-briefing.pdf (Accessed: 30 October 2020).

The group represented authors of leading academic papers on algorithmic audits and techniques, researchers in mis- and disinformation, policymakers in relevant regulatory bodies and policy departments in the UK and Europe, industry practitioners in data science, and civil society organisations with expertise in algorithm transparency, accountability and societal impacts of technology.

Background: COVID-19 misinformation

The COVID-19 crisis has shone a spotlight on well-established problems in digital society, with mis- and disinformation cross-cutting a range of online harms. In response to the crisis, technology companies have taken unprecedented steps to counter false and misleading content.

In March 2020, seven major companies published a joint statement on their commitment to combat misinformation, promote authoritative material and keep their communities safe.[7]

Facebook's response included banning ads intended to create panic, and the removal of false, potentially harmful claims as identified by health organisations. WhatsApp launched a partnership with the World Health Organisation (WHO) to provide COVID-19 updates, and introduced message-forwarding limits. Twitter expanded its definition of harm on the platform (used for content moderation and to identify violations of the platform), and YouTube featured verified COVID-19 content on its homepage.[8]

Automation has been a significant feature of companies' COVID-19 response, with companies scaling their reliance on machine learning to identify, triage or remove harmful content.[9, 10]

Despite these efforts, polling by the UK regulator Ofcom showed worryingly high exposure to disinformation; in May 2020, 50% of respondents said they encountered false or misleading information on a weekly basis.[11] Research by the Institute for Strategic Dialogue (ISD) further documented concerning gaps between stated policies and outcomes.[12]

[7] Microsoft. (2020) 'A joint industry statement on COVID-19 from Microsoft, Facebook, Google, LinkedIn, Reddit, Twitter and YouTube'. *Twitter.* Available at: https://twitter.com/Microsoft/status/1239703041109942272/photo/1 (Accessed: 30 October 2020).

[8] C. Colliver and J. King. (2020) 'The first 100 days: coronavirus and crisis management on social media platforms'. *Institute for Strategic Dialogue*. Available at: https://www.isdglobal.org/wp-content/uploads/2020/06/20200515-ISDG-100-days-Briefing-V5.pdf (Accessed: 30 October 2020).

[9] Sumbaly, R. et al. (2020) 'Using AI to detect COVID-19 misinformation and exploitative content'. *Facebook AI blog*. Available at: https://ai.facebook.com/blog/using-ai-to-detect-covid-19-misinformation-and-exploitative-content (Accessed: 30 October 2020).

[10] Derellla, M. (2020) 'An update on our continuity strategy during COVID-19'. *Twitter blog.* Available at: https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html (Accessed: 30 October 2020).

[11] Ofcom. (2020) 'Covid-19 news and information: consumption, attitudes and behaviour,' *Ofcom.* Available at: https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/news-media/coronavirus-news-consumption-attitudes-behaviour/interactive-data (Accessed: 30 October 2020).

[12] C. Colliver and J. King. (2020) 'The first 100 days: coronavirus and crisis management on social media platforms'. *Institute for Strategic Dialogue*. Available at: https://www.isdglobal.org/wp-content/uploads/2020/06/20200515-ISDG-100-days-Briefing-V5.pdf

The ISD and the BBC found that websites known to host disinformation about coronavirus had received over 80 million interactions on public Facebook pages since the start of the year.

In the same period, links to the US Centers for Disease Control and Prevention (CDC) and WHO websites gathered around 12 million interactions combined. Similarly, a study by Avaaz found that 100 pieces of coronavirus-related misinformation had been shared over 1.7 million times on Facebook and viewed an estimated 117 million times. 41% of the misinformation Avaaz analysed had remained on the platform without warning labels, although 65% had been debunked by partners of Facebook's own fact-checking program.

With this context, workshop participants considered algorithm inspection based on the following two scenarios:

Scenario 1: a content recommendation algorithm amplifies COVID-19 misinformation

Scenario 2: a content moderation algorithm fails to sufficiently identify and mitigate COVID-19 misinformation

As previously recommended by the Ada Lovelace Institute, policymakers will need to address gaps in the proposed regulator's legal authority and powers, organisational capacity and relevant skillset if they are to conduct a robust algorithm inspection.[13] The Ada Lovelace Institute and Reset convened this workshop to expand on these requirements.

[13] Ada Lovelace Institute and DataKind UK. (2020). 'Examining the black box: tools for assessing algorithmic systems'. *Ada Lovelace Institute*. Available at: https://www.adalovelaceinstitute.org/wp-content/uploads/2020/04/Ada-Lovelace-Institute-DataKind-UK-Examining-the-Black-Box-Report-2020.pdf (Accessed: 30 October 2020).

Insights

Insight 1: The current model of self-regulation is insufficient and cements information asymmetry between social media platforms and the public

At present, social media platforms apply voluntary standards. Each platform develops their own policies with minimal statutory or regulatory obligation, and releases transparency reports at their discretion. The platforms 'hold all the cards: they draw up their terms of use, decide to what extent to be bound by them, modify them as necessary without any public formalities'.[14] This creates significant information asymmetry; public authorities and other relevant third parties cannot access the evidence required to analyse harms related to the platform, or to assess the efficacy of platforms' policies and interventions.[15]

Currently, most independent third parties studying algorithmic impact rely on methods that require limited technology company involvement. Often the evidence they're seeking is not available at all, or relies on adversarial external audit methodologies that can be perceived as breaching a platform's terms of service.

External audit methodologies currently in use include:

- **Surveys or polling of users**. The UK regulator Ofcom's rolling survey of COVID-19 news and information consumption found that social media is consistently reported as the biggest source of misinformation, with 'theories linking the origins or causes of COVID-19 to 5G technology' being the most common misinformation.[16] Surveys or polling could also be applied to more targeted user bases of particular platforms under inspection.

[14] Direction interministérielle du numérique et du système d'information. (2019) 'Creating a French framework to make social media platforms more accountable: acting in France with a European vision'. *République Français*. Available at: https://minefi.hosting.augure.com/Augure_Minefi/r/ContenuEnLigne/Download?id=AE5B7ED5-2385-4749-9CE8-E4E1B36873E4&filename=Mission%20Re%CC%81gulation%20des%20re%CC%81seaux%20sociaux%20-ENG.pdf (Accessed: 30 October 2020).

[15] Center for Data Ethics and Innovation. (2020) 'Review of online targeting: final report and recommendations'. *Gov.uk*. Available at: https://www.gov.uk/government/publications/cdei-review-of-online-targeting/online-targeting-final-report-and-recommendations (Accessed: 30 October 2020).

[16] Ofcom. (2020) 'Covid-19 news and information: consumption, attitudes and behaviour,' *Ofcom*. Available at: https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/news-media/coronavirus-news-consumption-attitudes-behaviour/interactive-data (Accessed: 30 October 2020).

- **Scraping, sockpuppet audits and crowdsourced user auditing.** External audit techniques can include data scraping or data collection from a sample of users, either through the creation of sockpuppet user accounts or crowdsourcing data from current users (e.g. users can install a browser extension for data collection[17]). These techniques are often adversarial and face challenges to reach a large enough number of users, or legal concerns around violating a platform's terms of use.

  In the USA, researchers and journalists studying online discrimination have been wary of violating the Computer Fraud and Abuse Act (CFAA), which makes it a federal crime to access a computer in a manner that 'exceeds authorized access'.[18] The legality of this activity was clarified in Sandvig v. Barr, the recent USA federal ruling that determined that violating a website's terms of service does not violate the CFAA. While the ruling was a victory for algorithm transparency, it highlights the legal uncertainty of external investigation.[19]

- **Existing public API data or datasets.** Currently, public APIs (such as those offered by Twitter, YouTube, and Reddit) enable researchers to map the spread of information on the platform. However, API data and public datasets are released at the company's discretion, and the data available and frequency of release can easily change.[20]

These methods provide partial access to data that is often 'biased, incomplete, and subject to a range of awkward technical and contractual restrictions that impede its usefulness for empirical research'.[21] Using the example of COVID-19, we see that long-existing frustrations have resurfaced afresh.

In April 2020, 76 civil society organisations called on social media and content-sharing platforms to preserve data related to automated COVID-19 content moderation, provide data access to researchers and journalists (subject to privacy considerations), and to publish them in transparency reports.[22]

[17] Who Targets Me. 'Install the free Who Targets Me browser extension to track political ads'. Available at: https://whotargets.me/en/ (Accessed: 20 October 2020)

[18] United States District Court for the District of Columbia. (2020) 'Sandvig v. Barr – memorandum opinion'. *ACLU.* Available at: https://ww.aclu.org/sandvig-v-barrmemorandum-opinion (Accessed: 30 October 2020).

[19] Williams, J. (2018). 'D.C. court: accessing public information is not a computer crime'. *Electronic Frontier Foundation*. Available at: https://www.eff.org/deeplinks/2018/04/dc-court-accessing-public-information-notcomputer-crime (Accessed: 30 October 2020).

[20] Mozilla. (2019) 'Facebook's ad archive API is inadequate'. *The Mozilla Blog*. Available at: https://blog.mozilla.org/blog/2019/04/29/facebooks-ad-archive-api-is-inadequate/ (Accessed: 30 October 2020).

[21] C. Puschmann. (2019) 'An end to the wild west of social media research: a response to Axel Bruns'. *Information, Communication & Society*. 22(11). Available at: https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1646300 (Accessed: 30 October 2020).

[22] Llansó, E. (2020) 'COVID-19 content moderation research letter'. *Center for Democracy and Technology.* Available at: https://cdt.org/insights/covid-19-content-moderation-research-letter/ (Accessed: 30 October 2020).

Insight 2: An algorithm inspection will require detailed evidence on companies' policies, processes and outcomes, and new methods of access to evidence

Types of evidence:
Workshop participants identified evidence required for an algorithm inspection within three categories – policy, process and outcomes. Together, the evidence would be used to understand harms occurring on the platform, the platform's expected behaviour in response to harms, and to verify platform claims about its actions and the role of its algorithms in mitigating or increasing these harms.

1. **Policy:** what policies the platform has on harmful content, and content promotion and recommendation more broadly. This would include definitions and intended actions, seeking evidence such as:

   - What does the platform consider harmful content and why?
   - How does it intend to act on content identified as harmful?
   - What are stated criteria for intervention (e.g. moderation actions such as content removal, demotion or labelling)?
   - What are the platform's policies on promoting or recommending content?

2. **Process**: how the policy is enacted. This includes algorithmic systems used to enact policy, as well as human processes such as review, moderation or curation – and the intersection of the two:

   - What are the processes by which harmful content is flagged, reviewed and actioned?
   - How are moderation guidelines devised and updated, and by whom?
   - What languages and expertise areas are covered by those teams working on content policy design and enforcement?
   - What resources are earmarked by companies for this area of their work?

3. **Outcomes:** data on platform metrics, content and behaviour, to enable analysis as to the impact of policy, processes and organic platform activity:

   - Are there metrics on the identification of harmful content, and actions taken in response to harmful content?
   - What percentage of content removals receive human review, and how many users appealed the removal?

Means of access:

Workshop participants identified the following means to access this evidence, all of which require some degree of participation from technology companies.

Method	Examples	Benefits	Challenges
Documentation	Policy documentation, including definitions of misinformation or harmful content, related platform rules and actions, and reasoning behind them	Provides evidence of the company's (claimed) expected behaviour Enables initial scrutiny of policy stance	Without details of company processes and systems, risk of being a high-level understanding of policy intent (and not of realities on the platform)
	Process documentation, including instructions given to manual content moderators	Provides evidence of the company's (claimed) expected behaviour Enables initial scrutiny of process design	If made public, risks making it easier to 'game' moderation system
	Technical system documentation, including: • tools used to identify and moderate information • content recommendation and sharing systems	Provides evidence of the company's (claimed) expected behaviour Enables initial scrutiny of technical design	If made public, risks making it easier to 'game' moderation systems Concerns about intellectual property
Self-reported metrics	Self-reported metrics on misinformation and harmful content, such as: • model performance for recommender and moderation systems (including false positives and false negatives) • commercial data for promoted content that's later moderated • engagement metrics for content that's later moderated	Provides evidence of the extent to which company believes it is meeting standards	Lacks independent verification Platforms can selectively choose what to report
API access	Access to new or extended APIs for an inspector, such as access to live platform data	Enables real time/rolling scrutiny of a system's inputs and outputs to verify function and impact	Ongoing access must be agreed upon Companies could manipulate data available through the API
Dataset provision	Datasets shared with inspectors could include samples of moderated and unmoderated content and/or training data to develop moderation or recommendation models	Enables independent scrutiny of system, and provides inputs and outputs to verify function and impact	Datasets provide a snapshot of a single point in time - they may become out of date as user behaviour or system algorithms change Datasets may be selective Privacy concerns for users

Method	Examples	Benefits	Challenges
Interviews	Interviews with staff beyond the typical policy and legal teams who interface with regulators, such as: • Technical staff on product teams focused on moderation and recommendation software (product managers, engineers, data scientists) • Moderation teams implementing policies	Direct access to those who design and implement systems will more quickly reveal the principles underpinning the system, and design and engineering decisions and trade-offs	The power dynamic of employer-employee relationship may pressure interviewees Technical staff themselves may not fully understand algorithm behaviour and output
Code access	Access to code that underpins moderation or recommendation systems	Allows interrogation of algorithms and verification of system function	Code changes over time; access would need to be ongoing to be meaningful Security threat of ongoing access to systems Privacy concerns for users Understanding the code would require technical expertise (which may vary by platform). This would likely be slow and would benefit from support of engineers working at the social media platform Concerns about intellectual property
Inspector-set test results	A test or dataset for companies to run on their platforms (or for the inspector to run through a private API), in order to collect test results This could include benchmark datasets for different types of harms (which could be used to compare performance across platforms, or for a single platform over time)	Allows access to information and systems that are not public without direct access to systems	Results are not independently verifiable; concerns raised about reliability Hard to set universal tests for different platforms due to different content formats or processes, and it’s challenging to keep them up to date as platforms develop

Skills to identify, analyse and discuss evidence:
Algorithm inspection requires a multidisciplinary skillset, including technical and social science expertise to conduct or oversee the inspections, and the policy and communications skills to foster public engagement and dialogue.

Workshop participants generally agreed that the regulator would be a cornerstone of a strong inspection ecosystem where other actors – including independent investigators – continued to conduct audits and inspections, although independent inspection was not sufficient on its own. There was general consensus that the inspection would need to be conducted by the regulator or by a third party appointed or approved by the regulator, either through a new field of registered auditors (as in financial services) or one or more national bodies.

In their current composition, regulatory and national bodies may not necessarily have the employees with relevant skills for an algorithm inspection, as many of those individuals work in academia, policy or industry. The regulator would need to be able to access this expertise in-house, or through third-party bodies. Developing technical capacity within the regulator, as well as the capacity within technology companies to respond to access and disclosure requirements, would be a significant step towards creating inspection powers.

Insight 3: Algorithm inspection brings with it significant opportunity, but will require careful design to deliver on its potential.

Participants identified several areas that deserve particular scrutiny when designing the ideal regulatory inspection practice. These include:

- **Creating coherence** both with existing regulation, such as in broadcasting, data protection, trade secrets and cybersecurity, and across platforms that may have different behaviours or formats
- **Building regulatory capacity** that avoids overburdening small, specialised regulators (e.g. for harms) that lack capacity to have an impact
- **Establishing an independent review mechanism to 'regulate the regulators'** and prevent overreach, identify interference or influence, and verify quality of work
- **Developing appropriate penalties** to ensure compliance, ensuring sufficient repercussion given scale and power of companies involved
- **Managing dynamic systems over time:** regulation is often seen as a fixed snapshot, but assessments of social media platform algorithms cannot be static as the systems themselves are dynamic
- **Confluence of algorithms:** major platforms deploy distinct algorithms that can interact in surprising ways. A regulator may not be able to inspect a single algorithm in isolation, which adds complexity to the inspection process and resource allocation, as well as to establishing mechanisms that prevent overreach.

Perceived risks that could be raised of a regulatory inspection regime include:

- **Regulatory capture by big tech companies.** Risk of outsized influence of larger tech companies over regulatory decisions, or influence or lobbying of the government in turn influencing the regulator.
- **Regulatory misuse for political reasons or suppression.** Risk that regulation may be used for political motives, rather than objective analysis and enforcement. In particular, risk that regulatory powers imagined and developed in democratic regimes to increase transparency and reduce harms may enable suppression of freedom of expression, privacy and human rights infringements in other regimes (either geographical or temporal).
- **Enabling bad actors.** Risk that exposing moderation mechanisms may let bad actors 'game' these mechanisms. This was perceived as a risk only for inspection regimes that share evidence publicly, beyond the regulator and vetted third-parties.
- **Overburdening small operators.** Risk that small or new social media platforms are unable to grow or thrive due to the capacity burden created by regulatory requirements in place. This would not be a risk if the regime were limited to dominant companies.

The workshop did not explore these areas in detail, or the viable solutions to these perceived risks; instead, it focused on the preconditions (including foundational regulatory infrastructure) that are critical to regulatory algorithm inspection.

Recommendations

1. The regulator must have compulsory audit and inspection powers

An independent regulator should be empowered and resourced to enforce platforms' due diligence and transparency obligations. This governance framework can only work on one condition: it requires transparency from the platforms to an independent regulator. The regulator should have the power to demand any type of granular evidence that is necessary for it to fulfil its supervisory tasks, and to impose fines or other corrective actions when platforms do not provide that information in a timely manner.

While regulatory inspection of algorithms may be new, there are regulatory analogies in industries as varied as financial compliance, food safety and pharmaceuticals.[23] As one example, financial regulation sets clear precedent for independent audits of large businesses with commercially sensitive data; this is now a commonplace standard that operates together with public reporting requirements.[24, 25]

In addition, there is relevant precedent for the oversight of large technology companies; these are useful even if the oversight mechanisms do not satisfy all the requirements set out in this paper. In the USA, the Federal Trade Commission's consent order – settled with Facebook in 2011 – required that Facebook submit to external auditing of its privacy policies and practices.[26] In the UK, the ICO can undertake audits to assess how data controllers or processors are complying with good practice in the processing of personal data. If necessary, the ICO can seek a warrant to enter, search, inspect and operate any equipment.[27]

With compulsory audit and inspection powers, a regulator would correct the information asymmetry that currently defines the public's relationship with large technology companies. These powers are essential for effective oversight and compliance; without them, a regulator would struggle to achieve its statutory goals.

[23] Ghosh, D. and Scott, B. (2018) 'Digital deceit II. a policy agenda to fight disinformation on the internet'. *Harvard Kennedy School Shorenstein Center on Media, Politics and Public Policy*. Available at: https://d1y8sb8igg2f8e.cloudfront.net/documents/Digital_Deceit_2_Final.pdf (Accessed: 30 October 2020).

[24] Beverton-Palmer, M. and Beacon, R. (2020) 'Online harms: bring in the auditors'. *Tony Blair Institute for Global Change*. 30 July 2020. Available at: https://institute.global/policy/online-harms-bring-auditors (Accessed: 30 October 2020).

[25] Beverton-Palmer, M. and Beacon, R. (2020) 'Analysis: applying the principles of audit to online harms regulation'. *Tony Blair Institute for Global Change*. Available at: https://institute.global/policy/analysis-applying-principles-audit-online-harms-regulation (Accessed: 30 October 2020).

[26] Ghosh, D. and Scott, B. (2018) 'Digital deceit II. a policy agenda to fight disinformation on the internet'. *Harvard Kennedy School Shorenstein Center on Media, Politics and Public Policy*. Available at: https://d1y8sb8igg2f8e.cloudfront.net/documents/Digital_Deceit_2_Final.pdf (Accessed: 30 October 2020).

[27] Demos, Doteveryone, Global Partners Digital, Institute for Strategic Dialogue, Open Rights Group. (2020) 'Algorithm inspection and regulatory access'. *Demos*. Available at: https://demos.co.uk/wp-content/uploads/2020/04/Algo-inspection-briefing.pdf (Accessed: 30 October 2020).Demos, Doteveryone, Global Partners Digital, Institute for Strategic Dialogue, Open Rights Group. (2020) 'Algorithm inspection and regulatory access'. *Demos*. Available at: https://demos.co.uk/wp-content/uploads/2020/04/Algo-inspection-briefing.pdf (Accessed: 30 October 2020).'.

2. The regulator's information-gathering powers must extend to evidence on policy, process and outcomes

To fulfill its oversight function, the regulator will need the legal authority to access all necessary evidence. The UK Online Harms White Paper proposes information-gathering powers, including the 'power to request explanations about the way algorithms operate'.[28] Beyond these explanations, a robust inspection will require information on company policy (what is the policy, and what are its goals?), process (how is the policy implemented?), and the data that supports it (how is the policy monitored, what metrics are used to track it and the data to verify those metrics?).

The regulator must have the authority to request evidence, and technology companies will need to develop corresponding capacity to respond to these requests, which could include methods such as interviews, API access, or disclosure of internal policy documentation.

3. Powers to access and engage third-party expertise

An algorithm inspection requires a multidisciplinary skillset, although relevant expertise for any given inspection will vary based on context and industry. While the regulator should have some skills in-house, it will need the ability to access and instruct third-party expertise. This could be through powers similar to those of the UK's Financial Conduct Authority, who can require reports from third parties, or through a new field of registered auditors. Alternatively, the regulator could give independent experts secure access to platform data to undertake audits on its behalf.

As recommended by the UK Centre for Data Ethics and Innovation (CDEI),[29] academics should be able to access certain datasets when studying issues of public interest. The regulator should have the powers to mandate this access, especially on issues such as disinformation, where independent research will be crucial to developing future public policy.

[28] UK Department for Digital, Culture, Media & Sport and UK Home Office. (2020) 'Online harms white paper'. Available at: https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper (Accessed: 30 October 2020).

[29] Center for Data Ethics and Innovation. (2020) 'Review of online targeting: final report and recommendations'. *Gov.uk.* Available at: https://www.gov.uk/government/publications/cdei-review-of-online-targeting/online-targeting-final-report-and-recommendations (Accessed: 30 October 2020).

Conclusion: towards public oversight of algorithms

Inspection of algorithms will prove essential to any regulator's toolkit – it will be impossible to provide effective oversight without it.

While regulatory algorithm inspections have yet to be conducted in practice, it is possible to draw insights from external algorithm investigations and audits, as well as from regulatory regimes in other sectors to ensure regulation keeps pace with the scale and speed at which algorithms are being deployed.

Social media platforms should be subject to more public oversight, especially given the fundamental role they play in a functioning democracy and society.

Governments must develop and enact a public policy agenda that regulates the digital marketplace, and aligns its interests with those of democratic and social integrity.

At the same time, we must customise audit regimes to be proportional to the types of companies under review, and anticipate and mitigate their associated risks, including the potential for abuse.

Ultimately, we need novel and innovative forms of governance to address these challenges.

About the Ada Lovelace Institute

The Ada Lovelace Institute is a research institute and deliberative body, established by the Nuffield Foundation in early 2018, in collaboration with the Alan Turing Institute, the Royal Society, the British Academy, the Royal Statistical Society, the Wellcome Trust, Luminate, techUK and the Nuffield Council on Bioethics.

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

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

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

We are funded by the Nuffield Foundation, an independent charitable trust with a mission to advance social well-being. The Foundation funds research that informs social policy, primarily in education, welfare and justice. It also provides opportunities for young people to develop skills and confidence in STEM and research. In addition to the Ada Lovelace Institute, the Foundation is also the founder and co-funder of the Nuffield Council on Bioethics and the Nuffield Family Justice Observatory.

About Reset

With our partners, Reset works to ensure that the commercial interests of Big Tech are compatible with the values of robust and resilient democracies.

Reset was launched in March 2020 by Luminate in partnership with the Sandler Foundation. Reset seeks to improve the way in which digital information markets are governed, regulated and ultimately how they serve the public. The far-reaching objective of this work will be to change how the Internet enables the spread of news and information, restoring its ability to serve the public interest and democracy over corporate profits and exploitative political interests. We will do this through new public policy across a variety of areas – including data privacy, competition, elections, content moderation, security, taxation and education.

To achieve our mission, we make contracts and grants to accelerate activity in countries where specific opportunities for change arise. We hope to develop and support a network of partners that will inform the public and advocate for policy change. We are already working with a wide variety of organisations in government, philanthropy, civil society, industry and academia.

— www.reset.tech
— @resetdottech
— hello@reset.tech

Ada Lovelace Institute
adalovelaceinstitute.org

Part of the Nuffield Foundation
Registered charity 206601