# Making Audits Meaningful

Overseeing the Use of AI in Content Moderation

AUTHORS

**Hannah Bloch-Wehba**, Texas A&M University, School of Law, USA
**Angelica Fernandez**, University of Luxembourg, Luxembourg
**David Morar**, George Washington University, Elliott School of International Affairs, USA

This policy brief was formulated within the framework of the Research Sprint on *AI and Platform Governance* organized by the Alexander von Humboldt Institute for Internet and Society (HIIG) Berlin, Germany (August-October 2020). All authors contributed equally to the formulation of the policy brief.

## Executive summary

While platforms use increasingly sophisticated technology to make content-related decisions that affect public discourse, firms are tight-lipped about exactly how the technologies of content moderation function. The laconic nature of industry disclosure relating to their use of algorithmic content moderation is thoroughly unacceptable, considering that regulators need to understand the platform ecosystem in order to design evidence-based regulations and monitor risks associated with the use of AI in content moderation. This white paper sets out to explain how and why audits, a specific type of transparency measure, should be mandated by law within the four clear principles of independence, access, publicity, and resources. We go on to unpack the types of transparency, and then contextualize audits in this framework while also describing risks and benefits. The white paper concludes with the explanation of the four principles, as they are derived from the previous sections.

## Introduction

Our reality straddles the physical and digital worlds, with so much of our interactions, and increasingly our lives, relying on virtual platforms to communicate, network, and even work. The way these platforms make sense of the hundreds of billions of pieces of content residing on their servers thus becomes a crucial and critical aspect of inquiry. Relying on human action and reaction on such a large scale to moderate content is not realistic, and these platforms are now relying on algorithmic content moderation (ACM). Gorwa, Binns and Katzenbach define algorithmic (commercial) content moderation "as systems that classify user-generated content based on either matching or prediction, leading to a decision and governance outcome."[1] The increasing reliance on such tools is not in itself the issue as much as making sure that the algorithms and their use are accountable and transparent.

Unfortunately, companies have not been forthcoming about this very issue, which makes the need for accountability and transparency even more acute. A dearth of information is not an acceptable status quo, when citizens would be best served understanding the way a platform works in order to make informed decisions on whether and how to engage online. Even more, the laconic nature of industry disclosure relating to their use of ACM is thoroughly unacceptable, considering that regulators need to understand the platform ecosystem in order to design evidence-based regulations and monitor risks associated with the use of AI in content moderation. In spite of this, or perhaps because of it, these same regulators are considering going down the path of information-forcing or transparency-oriented rules.

This white paper sets out to explain how audits, a specific type of transparency measure, should be mandated by law within the four clear principles of independence, access, publicity, and expertise. We go on to unpack the types of transparency, and then contextualize audits in this framework while also describing risks and benefits. The white paper concludes with the explanation of the four principles, as they are derived from the previous sections.

## Forms of Transparency

Enhancing algorithmic transparency in the context of content moderation has been a significant challenge for regulators. In the past years, different ways have been tried to enhance systemic transparency from platforms. From transparency reports submitted to regulators voluntarily to data access regime initiatives, platforms have engaged in providing more transparency to regulators. Together, all of these tools increase algorithmic transparency. Audits are one more tool for regulators to achieve this goal. However, since there are many misconceptions of what an audit is, and since this is potentially confusing for policymakers, we start by contextualizing audits within the variety of available options to regulators to enhance algorithmic transparency.

#### Audits

Defined literally, an "audit" means "a formal examination of an organization's or individual's accounts or financial situation."[2] But what exactly does it mean to "audit" an algorithm? An algorithmic audit offers a concrete mechanism for assessing whether automated processes comply with the law, **internal** company rules, and/or regulations. Audits may be either internal (conducted by a company itself) or **external** (conducted by a third party outside the company). The auditing practice advocated by this policy brief, that is as a tool for regulatory oversight, is distinct from 'scraping' practices sometimes referred to as "independent auditing" used by advocacy groups and researchers as a way of obtaining large-scale datasets about platform operations. To be fully effective, auditors must have access to not just the model

[1] Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 2053951719897945.

[2] Merriam-Webster Dictionary, Audit: https://www.merriam-webster.com/dictionary/audit

architecture, data sets, or source code, but also information about internal stakeholders, the design and development process, performance characteristics, embedded assumptions, or information about the effects of the algorithm. The idea of an audit also conveys a degree of formality and rigor. Audits are more than a haphazard or occasional inspection: the formality of the process, and the rules by which it proceeds, lend audits credibility with companies and the public alike. Unlike impact assessments, which usually occur before a new technology is adopted, audits can occur at any time (indeed, repeatedly).[3]

Scholars and advocates concerned about the growing unaccountable power of algorithms in public life have suggested that audits may provide an answer to understanding whether, for example, automated decisions perpetuate bias or discrimination.[4] The chief advocates for algorithmic audits have often been researchers interested in studying the ramifications of algorithmic governance for the public interest. The core problem with audits is access to information: platforms often insist on concealing algorithms in order to protect either trade secrets, the integrity of automated decisions, or other interests. These interests in confidentiality and secrecy have hampered would-be auditors' access to key documentation, personnel, and systems. In light of this recalcitrance, researchers have struggled to find ways to audit algorithms, often turning to reverse-engineering, testing, and otherwise finding ways around unavailable code.[5]

Audits can also be useful investigative tools for regulators to gather information about compliance with legal and regulatory obligations. Just as tax audits allow tax officials to gather data about the subject of the audit, algorithmic audits can allow regulators to similarly gather data about how automated decision systems work within the social, technical, and organizational context of platform firms. And while researchers have to beg, plead, and grovel for access to proprietary information essential to auditors, regulators can require platforms to submit to audits and to make documentation, code, datasets, and other information available to auditors. Indeed, European law already contemplates the usefulness of audits in a closely related context: the General Data Protection Regulation (GDPR) ensures that supervisory authorities have the power to "carry out investigations in the form of data protection audits."[6]

Using algorithmic audits to investigate algorithms necessarily requires technical knowledge. And different technical designs will confront distinctive challenges to auditing. The design of a hash-based technology, which automatically screens user-generated content against a database of authoritative, "fingerprinted" content, is different from an artificial-intelligence technique, which might attempt to decide based on content and context whether user-generated posts are lawful or not.[7]

Algorithmic audits are also distinct from other forms of transparency, detailed below. Most importantly, audits facilitate direct access by regulators or their designated representatives to the mechanisms and systems of ACM. In this respect, audits have distinct advantages for the purposes of empowering regulators and informing regulatory oversight efforts. Other transparency requirements may serve other important interests, including individual interests in understanding how platforms reach content-related

---

[3] IEEE Standard 1028-2008 (defining audits as conducted by third parties); Inioluwa Deborah Raji et al., Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing, arXiv:2001.00973 [cs] (2020) (arguing for internal as well as external audits).

[4] Sandvig et al., An Algorithm Audit, in S.P. Gangadharan, V. Eubanks, & S. Barocas (eds.), Data and Discrimination: Selected Essays, https://www.newamerica.org/oti/policy-papers/data-and-discrimination/.

[5] Christian Sandvig et al., Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms 23 (2014); Jenna Burrell, How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms, 3 Big Data & Society 2053951715622512 (2016).

[6] GDPR art. 58.1.b; Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (2016) (calling for audits); Bryan Casey et al., Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise, 34 Berkeley Tech. L.J. 143 (2019).

[7] One of the authors has explained the details of these distinctions in prior work. Hannah Bloch-Wehba, Automation in Moderation, Cornell Int'l L.J. (forthcoming 2020).

decisions, but they are substantially less effective at informing regulatory efforts to an appropriate degree of granularity. Importantly, adopting algorithmic audit requirements need not be in conflict with these other transparency requirements. Indeed, audits can comfortably sit alongside notice, transparency reports, data access regimes, and registers as another way of promoting transparency, particularly with an eye toward effective regulatory oversight.

**Notice**

Rather than facilitating informed and effective regulation, much of the scholarly and policy debate about algorithmic transparency has focused on rendering platform decision-making legible to the individual users who are directly affected. Notices are the standards for explaining content moderation decisions by platforms. They inform users what action has triggered a moderation decision or an account suspension by the platform. However, notices are voluntary and often lack enough information for users to understand the decision that has been taken by the platform. Under the Santa Clara Principles, researchers agreed that notices as transparency mechanisms still need to be reviewed by platforms to include additional elements of information to be truly useful and provide a legal basis for accountability.[8]

**Transparency reports/databases**

Moreover, in recent years and to tackle harmful content, transparency reports by platforms have become a soft-law standard. Obligations to produce transparency reports are embedded in the EU Code of Conduct on Disinformation, the EU Code of Conduct on countering illegal hate speech, and in the EU draft regulation for countering online terrorism content. This type of transparency measure has also been included in different national legal instruments in Germany, Brazil , and Turkey.[9] However, regulatory authorities in the EU, such as the European Regulators Group for Audiovisual Media, have assessed transparency reports as an insufficient measure due to the lack of comparable information between the different platforms that have each their own metric and standards.[10] Moreover, they have criticized the lack of critical variables in these types of reports for regulators to monitor risks in the use of automated systems adequately.

**Data access regimes**

Data access regimes have also emerged as a way to increase transparency in platform governance. The most famous examples are Social Science One, Stanford University's Internet Observatory, and Microsoft Research Open Data program.[11] Platforms can also enter into specific data-sharing arrangements with governments or researchers. While these measures help researchers and policymakers to understand better automated systems, their voluntary nature is presented as challenging for designing an adequate

[8]On notices in the copyright context see: Perel, Maayan, and Niva Elkin-Koren Accountability in Algorithmic Copyright Enforcement. 473 SSRN (2015). On how notices on social media platform do not necessarily increase transparency for users, see: Nicolas P. Suzor et al., What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation, 13 International Journal of Communication 18 (2019).

[9] Some of the examples of national legal instruments that have embedded transparency reports obligations are: 1) NetzDG ( Art. 2) https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html; 2) Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet (Section II, Art. 6) https://edemocracia.camara.leg.br/wikilegis/p/12-lei-brasileira-de-liberdade-responsabilidade-e-transparencia-na-internet/.

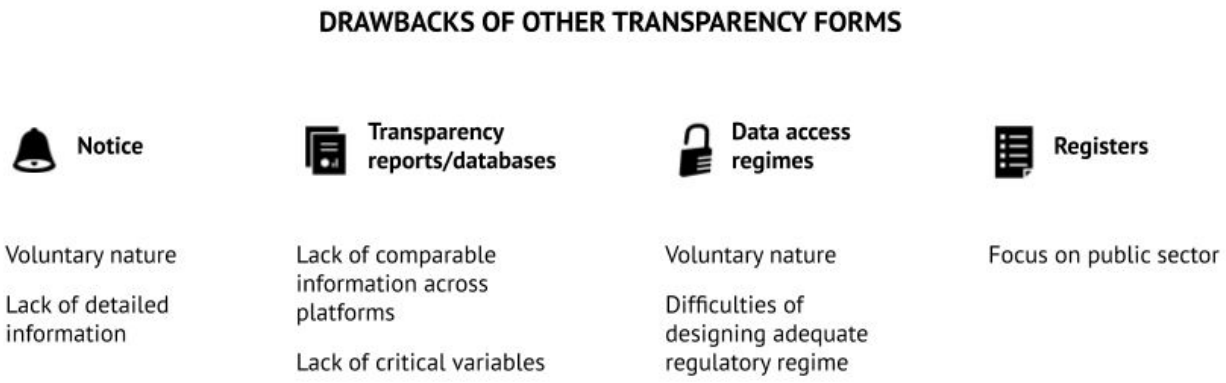
[10] See ERGA Report on disinformation: Assessment of the implementation of the Code of Practice at https://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf

[11] For more information on each of the data access regimes: a) Social Science One at https://socialscience.one/; b) Stanford University's Internet Observatory at https://cyber.fsi.stanford.edu/io/io; ; and c) Microsoft Research Open Data program at https://www.microsoft.com/en-us/research/project/microsoft-research-open-data/.

regulatory oversight of automated systems. Moreover, getting access to data solves only one part of the puzzle for regulators.[12]

**Registers**

Finally, within the context of AI systems, we have several initiatives of algorithm registers. These registers are being proposed as a way of achieving a baseline level of transparency. This type of transparency measure has been raised internationally and is now being tested in the European cities of Amsterdam, Helsinki, and Nantes.[13] Their main objective is to institute public registers as a mechanism for mandatory reporting of automated decision-making systems. Although these registers are aimed at enhancing transparency in the public sector, it is possible to imagine that they could also be used by private companies to better inform their users on the type of ADM systems used.




To conclude, all of these measures enhance transparency in different ways, but they are not to be confused with audits. Unlike audits, none of these measures promote direct regulatory access to critical proprietary information held within firms. Without more robust measures to facilitate regulatory oversight, neither transparency toward users nor more programmatic data-sharing or transparency initiatives will be sufficient to ensure that the regulatory state can keep up with innovations in the private-sector.

## Benefits and Risks of Auditing ACM

**Key Benefits**

One of the main benefits of audits for regulators is that they provide material information for regulators and the public on how an AI system works. This is essential to enable the verifiability and explainability of these systems.

[12] Mike Ananny & Kate Crawford, Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability, 20 New Media & Society 973 (SAGE Publications 2018).

[13] Amsterdam city AI register https://algoritmeregister.amsterdam.nl/en/ai-register/; City of Helsinki https://ai.hel.fi/en/ai-register/ ; Nantes Metropole AI register https://data.nantesmetropole.fr/pages/algorithmes_nantes_metropole/

From a technical point of view, algorithms can malfunction. Computer scientists and researchers[14] have identified several sources of error. It could be a mathematical problem in the code, and therefore, the algorithm does not solve the problem correctly or completely or it could be an incorrect modeling or implementation of an algorithm. In the case of learning algorithms, the source of error could be the unsuitability or faulty training data used by the system. Finally, a correct algorithm can still have socially undesirable side effects when it interacts with human behavior. The only way for regulators to verify the correctness of the implementation of an algorithm is by having access to the code and understanding the underlying mathematical problem that the algorithm tries to solve as well as the assumptions made when translating the question into a mathematical formula. Therefore, only disclosing the code to regulators or academics is not enough to verify an automated system since the code by itself without explanation does not allow regulators to assess compliance and make recommendations on how to improve.[15]

A crucial benefit of implementing algorithmic audits is that it allows regulators to go beyond risk-based assessments frameworks implemented by companies as a way of compliance. A risk-based approach often overlooks social and ethical challenges and only focuses on probabilities. Further, audits under this framework are traditionally done as an ex post examination. However, in the context of AI systems, researchers consider that audits could be used as a mechanism to check ex ante that the engineering processes involved in AI systems' design and implementation are in order with ethical expectations and standards such as AI principles or guidelines.[16] In most industries we understand audits as being performed as an ex post meticulous and methodical analysis, but using audits to anticipate potential system-level risks, and design adequate monitoring strategies for potential adverse outcomes is an advantage for regulators. Also, if mandated, audits will provide a way of operationalizing ethical principles or guidelines from the starting point of software development.

Finally, audits provide a transparency trail for regulators. One systematic challenge for AI researchers is to trace the provenance of training data or to interpret the meaning of model weights, which influence the risk profile of an AI system. Documenting the relationship between product requirements, their source, and system design helps to close the gap on accountability on the use of AI systems. Traceability is an essential concept already used in the standard for auditing software.[17] This notion has been transposed to audits in different domains and has proven as an indispensable notion when transposed to high risk domains, such as the aerospace industry. It has also been identified as a key for enabling algorithmic accountability.

The benefits of opting for auditing has not been lost on large platforms, such as Facebook and Google,[18] who have been reported to be looking into internal auditing practices. However, our proposal focuses on developing a framework for national or supranational regulators to undertake this task, and not solely as an internal obligation for platforms.

---

[14] Algorithm Watch. Working paper: Verifiability of algorithms at https://algorithmwatch.org/publication/zweites-arbeitspapier-ueberpruefbarkeit-algorithmen/.

[15] Andrew D Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. Fordham L. Rev. 87 (2018), 1085.

[16] Inioluwa Deborah Raji et al., Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing, arXiv:2001.00973 (2020)

[17] IEEE. 2008. IEEE Standard for Software Reviews and Audits. IEEE Std 1028-2008 (Aug 2008), 1–53. https://doi.org/10.1109/IEEESTD.2008.4601584

[18] Johnson, K. (2020). Venture Beat. Retrieved from https://venturebeat.com/2020/01/30/google-researchers-release-audit-framework-to-close-ai-accountability-gap/

Finally, audits allow operationalizing ethical principles into practice for platforms using AI systems. Given the multitude of ethical principles for AI which are under discussions at international and national level, audits come as a tangible tool for regulators to design an independent oversight framework.

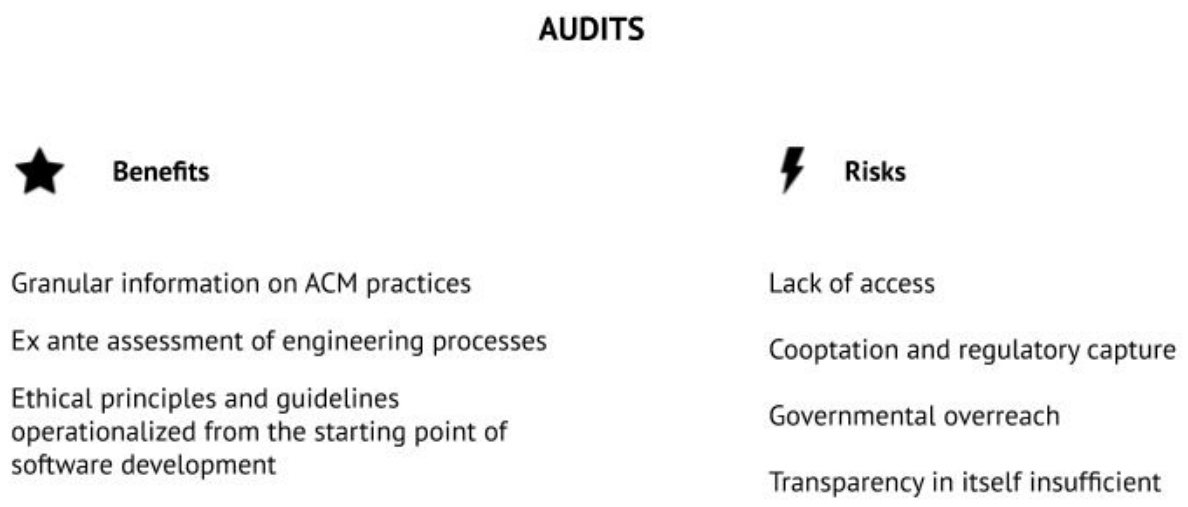
**Potential Risks**

Like any regulatory initiative, audits also have their drawbacks. Some auditing regimes rely heavily on third-party independent auditors, who sometimes lack adequate access or power to fully inform the public about company practices. For instance, in 2012, the United States Federal Trade Commission issued an enforcement order requiring Facebook to submit to independent external audits of its mechanisms for overseeing the security practices of third-party app developers. But it's auditor, PriceWaterhouseCoopers, referred only to Facebook's publicly available policies in its report, casting doubt on its ability to fully assess Facebook's internal practices.

Calls for audits may also inadvertently lead to cooptation and watering-down of the term. Consider how a virtual cottage industry of "auditing" and compliance monitoring has arisen as a result of the passage of the GDPR. To the extent algorithmic audits are interpreted as nothing more than a box on a compliance checklist to be ticked off by a private certification body, they will fail to provide meaningful information or access to regulators or to the public. In other words, robust audits must not be turned into nothing more than a rubber stamp of self-certification. If regulatory mandates for auditing are not carefully and thoughtfully crafted, they are unlikely to generate information of use to the public or to regulators. But mandatory, public audit provisions, which empower regulatory bodies to require independent third-party audits of platforms, can yield important information for public oversight.

Regulatory oversight can also be leveraged in ways that are regressive. As nations experiment with novel ways of governing social media platforms, certain initiatives run the risk of infringing user privacy, stifling dissent, and otherwise weaponizing the regulatory process to dampen, rather than promote, fundamental rights. Consider, for example, Russia, where recently enacted regulations are intended to centralize and consolidate state control over Internet companies and traffic, permitting already-extensive state surveillance to grow unchecked. As with any regulatory initiative, audit requirements must be carefully engineered to guard against the risk of abuse and misuse, politicization, and overreach.

On an even broader level, a core question concerns the value of transparency itself in promoting regulatory accountability. Some critics have charged that algorithmic transparency itself can serve as a distraction from more pressing and fundamental accountability concerns. And "thin" versions of platform transparency requirements, while appealing to the private sector, may be insufficient to ensure an adequate flow of information to either users or to regulatory agencies. Consider the example of the voluntary Santa Clara Principles on Transparency and Accountability in Content Moderation, which call on platforms to provide more detailed aggregate data about content removal, to notify each affected user about the reason that a post was removed or account suspended, and to provide opportunities for users to appeal those decisions. While these principles are laudable, their focus on individual user rights has meant that they have scarcely affected regulatory oversight. The Principles thus illustrate how transparency initiatives intended to facilitate effective regulation must be designed differently than transparency initiatives primarily designed to inform the public.




## Principles

Our broad overview of transparency and audits makes clear that, unless properly constructed, audits themselves may end up being useless or, worse, harmful to the public interest. This should not be perceived as a critique of the concept itself, as much a constructive assessment of how to best shape such an important tool in better understanding ACM. Simply making audits legally mandated will not be enough to properly extract important information for future regulatory and user actions. Four important principles need to be satisfied so as to ensure a baseline usefulness for audits: independence, access, publicity, and funding/expertise.

#### Independence

Independence is a straightforward concept, but the devil is in the details. In the auditing of ACM, it implies the existence and active participation of a third-party that is carrying out the audit. The current environment where companies are unwilling to provide information about their practices indicates the need for such a crucial principle to be embedded. This principle is rooted in the perspective that capture must be avoided: the audits must not be tied to the whims of the company being audited, and it also should not be entirely beholden to the State. Culturally, societally, notions of government intervention are different around the world and suggesting a particular model may not be entirely useful to all regulators.

Models for accrediting third-party auditors may take several different forms. One option is to follow the European model of media regulation and task Independent State Authorities to accredit auditors or to carry out algorithmic audits themselves. A second possibility would be for the state to recognize and accredit self-regulatory bodies that may conduct the audits themselves, consistent with regulatory objectives. A third possibility is for multi-stakeholder institutions to create the frameworks and the implementation of accreditation procedures for third-party auditors. There are many equations that eventually reach the same finality of the strict but broad principle of independence. We believe that this principle is satisfied as long as the outcome is a process that is not easily captured by either government or industry, be it through an equal say for the two, or an elevation of civil society to decide with input from both the state and platforms.

**Access**

Access stems from the current system where companies pick and choose which information to share and to whom. While typical answers about trade secrets and protecting corporate interests are the usual roadblock, it is imperative that industry does not become the gatekeeper of information. There are certainly legitimate concerns about competitors stealing data or algorithms, and those should be acknowledged. But access does not imply a free-for-all open door policy where the world can just peek in at will. Nor does it mean a wanton desire to hurt industry under the guise of helping society. It simply implies that, much like independence, the sole decision-making power about whether regulators ought to be able to peek under the hood of content moderation techniques should not rest with companies. Tiered systems of access can provide an answer to these concerns while also allowing in the daylight needed to conduct proper review of the systems being audited.

**Publicity**

Publicity implies that the audits should be well publicized. While primarily a tool for regulators and policy-makers to better understand the specific algorithms in order to craft better regulations, audits are also of service to the population in order to make informed choices. The mere existence of information, however, is not enough to claim that citizens and regulators are making informed decisions. Hidden behind corporate websites, written in undecipherable technical or legal languages, housed under confusing menus is not an actual form of transparency, but one of obfuscation, infoglut, and deliberate acts of opacity. Audits should be legally mandated to be widely available, in multiple formats, with accessibility in mind. This can be done either individually by companies, or through a system of public registers. As with previous principles we do not believe that one answer should be the choice for all regulators, but publicity should be construed and institutionalized as more than just a genuine effort on behalf of the platforms. It should be an intentional form of reaching intended audiences, both regulators and the public, in an understandable and contextualized manner.

**Resource**

Resource is based on the clear idea that audits are resource-dependent. Platform-based funding for audits should be legally mandated, so that a lack of financial resources can never become the reason for transparency to falter. Internet companies using ACM spend, depending on their size, various fortunes on their systems, which would imply that making these algorithms successful is worth the investment. A societal shift currently taking place is moving the window of acceptability on what algorithms can and should be allowed to decide, as well as the ways in which they do it, which, in turn, is changing the definition of what it means for an algorithm to be successful. Funding audits should happen through foundations or institutions that are otherwise disconnected from industry, in order to ensure a firewall between the results of the audits and continued funding from companies. Such a scheme could work either as a pool of money, funded by mandatory fees based on the size of the company (size to be construed in any way that makes sense for all stakeholders) or based on direct payment for the audit, through an independent body.

Even more, beyond financial resources, audits, as made clear above, are dependent on legitimate technical expertise. This means that an understanding of the technologies, mechanisms, uses, and outcomes is necessary both in crafting the structure of the audits and in carrying them out. Technical and financial limits to resources should be alleviated through clear legislation. By requiring audits, our hope is that the regulatory state will quickly develop and foster the technical expertise both to understand algorithmic audits and to make use of the information they provide.

## Conclusion

This white paper sets out to identify and attenuate a fundamental concern related to how we communicate online. We argue that while not a concern in itself, the use of ACM in a secretive manner makes it difficult for the public and for policymakers to take action for the present (choice of use) and the future (choice of regulation), respectively. There are many ways to establish the practice of transparency, however our research points to audits as the most robust way of understanding ACM. While certainly not a silver bullet, audits can be crafted in a way that as best as possible assuages its empirically obvious limitations while enhancing its potential benefits. We propose that mandating audits to be independent, to have built-in access, to be widely publicized, and to be resource-sufficient, be it in funding or expertise, are the crucial ways in which audits as a tool can help lead us closer to better understanding the choices made in ACM.

# References

## Articles and Research Papers

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. N*ew Media & Society*, *20*(3), 973-989.

Bloch-Wehba, H. (2020). Automation in Moderation. *Cornell International Law Journal, Forthcoming*.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512.

Casey, B., Farhangi, A., & Vogl, R. (2019). Rethinking Explainable Machines: The GDPR's Right to Explanation Debate and the Rise of Algorithmic Audits in Enterprise. *Berkeley Tech. LJ*, *34*, 143.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 2053951719897945.

Johnson, K. (2020). Google researchers release audit framework to close AI accountability gap. *Venture Beat*. Retrieved from https://venturebeat.com/2020/01/30/google-researchers-release-audit-framework-to-close-ai-accountability-gap/

Pasquale, F. (2015). The secret algorithms that control money and information. Harvard University Press.

Perel, M., & Elkin-Koren, N. (2015). Accountability in algorithmic copyright enforcement. *Stan. Tech. L. Rev.*, *19*, 473.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33-44).

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, *22*.

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham L. Rev.*, *87*, 1085.

Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, *13*, 18.

## Policy documents, studies and contributions

Algorithm Watch. (2016). Überprüfbarkeit von Algorithmen. *algorithmwatch.org*. Retrieved from https://algorithmwatch.org/publication/zweites-arbeitspapier-ueberpruefbarkeit-algorithmen/.

ERGA (2020) ERGA Report on disinformation: Assessment of the implementation of the Code of Practice. Retrieved from https://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf

IEEE (2008). IEEE Standard for Software Reviews and Audits. IEEE Std 1028-2008 (Aug 2008), 1–53. https://doi.org/10.1109/IEEESTD.2008.4601584

Sandvig et al. (2014). An Algorithm Audit, in S.P. Gangadharan, V. Eubanks, & S. Barocas (eds.), Data and Discrimination: Selected Essays. Retrieved from https://www.newamerica.org/oti/policy-papers/data-and-discrimination/.

## Legislation

European Union (2016). General Data Protection Regulation (GDPR) [art. 58.1.b]

German Federal Ministry of Justice and Consumer Protection (2017). Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken [Netzwerksdurchsetzungsgesetz] [NetzDG]) (DEU)

Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet (2020). (Section II, Art. 6)