# Position: Machine Learning-powered Assessments of the EU Digital Services Act Aid Quantify Policy Impacts on Online Harms

**Eleonora Bonel** [*1] **Luca Nannini** [*2 3] **Davide Bassi** [2] **Michele Joshua Maggini** [2]

## Abstract

While machine learning shows promise in automated knowledge generation, current techniques such as large language models and micro-targeted influence operations can be exploited for harmful purposes like the proliferation of disinformation. The European Union's Digital Services Act (DSA) is an exemplary policy response addressing these harms generated by online platforms. In this regard, it necessitates a comprehensive evaluation of its impact on curbing the harmful downstream effects of these opaque practices. Despite their harmful applications, we argue that machine learning techniques offer immense, yet under-exploited, potential for unraveling the impacts of regulations like the DSA. Following an analysis that reveals possible limitations in the DSA's provisions, we call for resolute efforts to address methodological barriers around appropriate data access, isolating marginal regulatory effects, and facilitating generalization across different contexts. Given the identified advantages of data-driven approaches to regulatory delivery, we advocate for machine learning research to help quantify the policy impacts on online harms.

## 1. Introduction

As machine learning (ML) becomes increasingly embedded in important societal systems and automated knowledge generation, it requires careful governance to mitigate emerging threats. In particular, recent advances in generative modelling, micro-targeted influence campaigns and selective exposure pose risks of exacerbating disinformation spread if deployed irresponsibly by profit-driven platforms. For example, large language models (LLMs) exhibit capabilities to generate synthetic text and imagery that may promote false narratives (Da San Martino et al., 2021; Zhou et al., 2023). The computational propaganda tactics involving psychographic targeting that emerged in the Cambridge Analytica scandal revealed machine learning's risks in enabling manipulation at scale when coupled to unethical business practices (Bakir, 2020). Additionally, excessive personalization in recommender systems can distort exposure diversity, enabling 'filter bubbles' that intensify attitude polarization by limiting viewpoint plurality (De Biasio et al., 2023). These trends underscore the need to ensure transparency and oversight while keeping up with artificial intelligence (AI) developments. The latter, despite promising great progress, also carries potential for considerable societal harms from misuse.

In this context, an increasing body of policy responses aimed at managing and mitigating machine learning risks has emerged. The recent EU Digital Services Act (DSA) (European Parliament and Council, 2022a) represents one such policy effort to expand governance over major platforms accused of inadequately combating online harms. Effective November 2022, the DSA categorizes platforms by size and risk, imposing graduated obligations such as notice-and-takedown procedures, risk assessments, transparency duties, and independent auditing. It targets issues like election interference, public health emergencies, and the spread of disinformation by requiring platforms to exercise heightened due diligence.

Ultimately, as the expectations on the scope of such a regulatory text grow, evidence-based assessment of their real-world impacts becomes pivotal. Open questions remain around the DSA's tangible impacts amidst the legal, technical and enforcement ambiguities characterising dominant platforms' governance of algorithmic accountability and content moderation practices (Leiser, 2023; Klonick, 2017; Hacker et al., 2023; Heldt, 2022; Sullivan, 2023; Kuczerawy, 2021; Trust Lab, 2023). Indeed, without empirical assessment, the effectiveness of this regulatory mechanism in achieving stated aims of mitigating viral falsehoods and protecting democratic integrity is yet not granted.

[*]Equal contribution [1]École d'Affaires Publique, Sciences Po, Paris, France [2]Centro Singular de Investigación en Tecnoloxías Intelixentes da USC (CiTIUS), Santiago de Compostela, Spain [3]Minsait, Indra Sistemas SA, Madrid, Spain. Correspondence to: Eleonora Bonel <eleonora.bonel@sciencespo.fr>, Luca Nannini <l.nannini@usc.es>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

Extending beyond the contextual analysis of the DSA presented in our previous work (Nannini et al., 2024), **this position paper argues that advanced analytical techniques, such as natural language processing, causal auditing, and diffusion modelling, are essential yet underutilized tools for public oversight over ensemble systems enabled by modern machine learning**. After highlighting some harmful ML trends, and then identifying some limitations of the DSA's response in this regard, we propose a research agenda to strengthen proportionality and participatory oversight in this new area of socio-technical governance. Sustained, coordinated efforts to overcome data access barriers remain imperative between scientists, civil society and policymakers aiming to use analytical methods, in the context of DSA implementation. By contributing to an enhanced methodology, we seek to increase transparency and evidence-based policy-making without stifling continued innovation.

In the following sections, we initially provide a summary of key harmful ML trends in social media applications (Section 2) which, in turn, the DSA's core provisions seeks to address (Summarised in Section 3). Subsequently, we turn to the DSA's limitations, highlighting the background (Section 4) of machine learning techniques for impact analysis. We then propose an outline of priorities for advancing evidence-based assessment of the DSA by discussing methodological barriers (Section 5), and ultimately issuing a call (Section 6) for coordinated research efforts, with some concluding remarks (Section 7).

## 2. Concerning Influence of Current ML Trends

Recent advances in applied machine learning have delivered transformative capabilities for automated content generation, profiling and information filtering. These techniques also carry risks of exacerbating disinformation spread if misused by profit-oriented platforms. We highlight three key trends necessitating governance.

### 2.1. Generative Modeling Capacities

Generative AI models exhibit an ability to automatically generate synthetic text, imagery and audio that potentially furthers false narratives. The low cost and accessibility of textual generators could favor mass production of manipulative messaging, based on learning statistical associations from vast training data (Johnson et al., 2022; Morrison, 2023; Heikkilä, 2023; Nannini, 2023; Neff, 2024; Lopatto, 2024). Similarly, LLMs are still insufficiently informative and reliable when it comes to sensitive topics, such as political science or business scenarios (Romano et al., 2023; Notopoulos, 2023; Mok, 2023; Perrigo, 2023; Kelly, 2023; Kaye, 2023; Sankaran, 2023). However, determining the appropriate thresholds and processes for limiting generative speech poses complex ethical dilemmas, given the importance of open inquiry. Moreover, detecting increasingly sophisticated AI-generated synthetic media remains an arms race with consistently evolving generative capacities (Yu et al., 2021; Du et al., 2020; Mirsky & Lee, 2022). Recent models have demonstrated the ability to generate multi-paragraph content emulating human writing styles on arbitrary topics, with sufficient coherence to deceive readers (Brown et al., 2020; OpenAI, 2023; Spisak & Edunov, 2023; MistralAI, 2024). Generative image, video and audio present additional challenges as visual markers of integrity become manipulable (Khoo et al., 2022; Asnani et al., 2023). More concerningly, generators directly optimized to manipulate people by eliciting emotions could automate influence campaigns based solely on data patterns, without ethical constraints (Sharma et al., 2024).

### 2.2. Microtargeted Influence Campaigns

The Cambridge Analytica revelations underscored the risks of combining psychologically-informed messaging based on machine learning-enabled profile analytics, with highly targeted digital advertising architecture (Bakir, 2020; Hinds et al., 2020). This significant case highlighted how hyper-personalized targeting, based on analyzing trace data reflecting interests and vulnerabilities, allows tailored influence at scale. Opaque auction-based ad infrastructure partitions audiences into marketable segments auctioned for profit (Choi & Lim, 2020; Yan et al., 2009; Gharibshah et al., 2020). This combination of audience partitioning and targeting enables digital rhetoric tailored to manipulate people by triggering identity-protective cognition and confirmation bias using their own data patterns against them (Tian & Wang, 2023; Acemoglu & Ozdaglar, 2011). By exploiting automated A/B testing,[1] these models can further optimizes persuasive message variants (Matz, 2021; Quin et al., 2024). While advertising microtargeting aims to maximize engagement, it could be co-opted for covert manipulation on consequential issues (Matz, 2021; Pham et al., 2022; Mavriki & Karyda, 2020). Its capacity for obscured discrimination also warrants examination (Speicher et al., 2018).

### 2.3. Filter Bubbles and Selective Exposure

Finally, much focus surrounds the risks of 'filter bubbles' from personalized recommender systems distorting exposure diversity to align with inferred preferences (Bozdag & van den Hoven, 2015). This could enable 'radicalization pathways' through increasingly niche content (Haim et al.,

[1]Automated A/B testing refers to a procedure where multiple message variants are tested by showing them to randomized user groups and measuring resulting engagement or conversions to incrementally maximize effectiveness.

2018). Data-driven feedback loops tracking engagement compound this by optimizing to appeal to people's confirmation biases through inflammatory material aligned with their interests. Critics argue that excessive personalization limits pluralistic discourse (Helberger, 2019). There are open questions around actual selective exposure effects in practice (Schäfer, 2023; Borgesius et al., 2016; Liao & Fu, 2013), also in relation to generative AI models (Sharma et al., 2024). However, unchecked optimization risks creating fragmented and polarized discourse arenas that are detrimental to democratic society (Pérez-Escoda et al., 2023).

Overall, the three trends highlighted in this section remind us of the urgency to ensure that transparency and oversight mechanisms keep up with the pace of AI developments. We saw how machine learning, while carrying promises of efficiency and reducing organizational burdens, can also exacerbate information disorders. In this context, ML threats include generating synthetic media to manipulate emotions, conducting microtargeted influence campaigns, and creating filter bubbles through personalised recommender systems that foster selective exposure. In the next section, we examine the DSA as a policy response tackling these machine learning threats.

## 3. The DSA Response & Open Questions

The European Digital Services Act (DSA) represents one prominent attempt at expanded platform governance, providing different rules for various online players according to their role, size and impact in the online ecosystem. Adopted in 2022 and applicable since Februrary 2024, the DSA places new requirements on very large online platforms (VLOPs) and very large search engines (VLOSEs), incorporating new legal requirements and granting new rights for online users. The regulation will require heightened due diligence around issues like online disinformation through heightened transparency, appeal and redress mechanisms, systemic risk assessment and mitigation, and independent auditing duties (European Parliament and Council, 2022a).

It is important to note that while our analysis focuses primarily on the DSA's provisions related to disinformation, the regulation also covers a wider range of online harms, such as hate speech, terrorist content, child sexual abuse material, and the sale of counterfeit or dangerous products (Recital 12). These harms often intersect with and amplify the spread of disinformation, creating complex challenges for platform governance and content moderation (Nannini et al., 2024).

While most of its provisions are enforced, it is still early to evaluate the regulation's overall success, as the path for effective implementation is expected to face evolving hurdles. Indeed, for the purposes of our analysis, we examine identified legal and technical limitations within the DSA's provisions that potentially constrain its aims to foster greater accountability. Nonetheless, the DSA signals an important policy effort to constrain unchecked harms in previously unregulated technologies, simultaneously to other pieces of legislation such as the AI Act and the Regulation (EU) No 900/2024 on transparency and targeting of political advertising (European Parliament and Council, 2024; Novelli et al., 2024).

### 3.1. Overview of Core DSA Provisions

To begin our analysis, we provide an overview of major mechanisms and associated open questions identified in the research literature, that may impede the DSA's intended legislative impacts without iterative refinement and oversight. The section only refers to the selection of DSA provisions that would warrant ongoing refinement, grounded in emerging evidence.

**Notice-and-Action Requirements.** A core pillar across online intermediaries involves standardizing notice-and-action mechanisms for reporting and expediting removal of 'manifestly' illegal content originating from third parties. The DSA's Article 16 mandates accessible reporting channels for any individual or entity to notify and provide substantiated explanation to providers of hosting services of the presence of items of information that the entity would consider to be illegal content. Compliance timeframes are delineated based on notice source, with reports by authorities or vetted expert entities necessitating urgent redress compared to ordinary individuals (European Parliament and Council, 2022a).

Defining 'manifestly illegal' content, establishing credible reporting standards, and ensuring participatory oversight of expanded take-down protocols are critical to preventing overreach that could impact legal speech or activism. In this regard, the process risks privileging institutional definitions of harm over lived experience, especially if implemented without sufficient checks against inconsistent restrictions on diverse viewpoints.

Nevertheless, the lack of consistent, detailed take-down protocols risks generating uneven enforcement, allowing platforms latitude to arbitrarily restrict or ignore legal speech (Sullivan, 2023; Wolters & Gellert, 2023). Critics highlight how absence of safe harbor limits on proactive automated filtering enables intermediaries to bypass human oversight and over-remove disputed borderline content without oversight mechanisms for challenging errant interpretations of alleged illegality (Hacker et al., 2023; Laux et al., 2021). This compounds transparency concerns around covert visibility reductions already occurring via shadow banning or demotions (Leerssen, 2023).

**Systemic Risk Assessments & Mitigation.** Very large online platforms (VLOPs) surpassing 45 million EU-located monthly active users must conduct recurring risk assessments examining if their systems, algorithms, data practices, business incentives, or design choices undermine public interests across domains like election integrity, media freedom, and fundamental rights. Assessed issue areas encompass security, privacy, fairness, accountability, safety, and democratic participation. Under articles 34-37, upon uncovering risks, platforms must implement reasonable, effective mitigations respecting proportionality principles and evaluated through mandatory independent audits (European Parliament and Council, 2022a).

Beside that, the notion of 'systemic risk' lacks unambiguous technical and procedural specificity regarding assessment methodologies (Sullivan, 2023; Heldt, 2022). Open questions around optimal quantitative indicators and qualitative review processes–whether to prioritize model-driven statistical likelihoods or contextual probability against severity – may allow self-serving risk ratings. For example, viral false information exploiting 'engagement'-maximizing feeds (Lo & Wei, 2023), 'filter bubbles' limiting exposure diversity (Benkler et al., 2018), or outrage-inciting dark design patterns may be underexamined if platforms minimize culpability. Similarly, spillovers to private groups (e.g., closed Telegram channels) remain excluded from assessments focused narrowly on public posts (Hacker et al., 2023).

**Independent Auditing Requirements.** To facilitate external accountability, VLOPs must commission annual independent audits, conducted per detailed specifications, examining their compliance with transparency duties, risk analyses, content moderation, and advertising systems (European Parliament and Council, 2022a; European Commission, 2023b). Audit reports with methodologies, findings, and key recommendations must be published to enable civil society monitoring (Kirchner & Reuter, 2020). The Digital Services Coordinators appointed in Member State can request additional audits on issues of concern (Jaursch, 2023).

Yet scholars identify that auditors still lack investigative capacities and genuine independence from the powerful platforms they evaluate based on financial contracts (Wilman, 2022; Laux et al., 2021). Following the DSA's recital 92, audits emphasize procedural soundness rather than judging substantive systemic outcomes, proportionality, fairness, or societal impacts enabled through compliant processes (European Parliament and Council, 2022a). Nevertheless, how effectively occasional audits can spur continuous internal improvements remains debated given platforms' control over information flows to reviewers (Laux et al., 2021). The level of ongoing cooperation and transparency firms provide or conceal post-audit is crucial for genuinely enhancing accountability rather than performative compliance.

### 3.2. Shortcomings for Harmful ML Applications in Social Media

While the DSA introduces strong provisions to request platforms' accountability, we observe specific limitations regarding risks from machine learning applications related to the spread of disinformation. Notably, gaps persist around the oversight of generative synthetic media, transparency over microtargeted influence campaigns, uncontrolled filter bubbles and selective exposure effects, and engagement-maximizing outrage incentives. These limitations are particularly important to identify for vetted researchers who, under Article 40 of the DSA, will be able to request further data from VLOPs and VLOSEs to conduct research on systemic risks.

**Gaps Regarding Generative Models and Synthetic Media.** The DSA's recital 81 states that certain risks, including risks to the right to human dignity, freedom of expression and of information, and data protection, may arise in relation to the design of algorithmic systems used by VLOPs or VLOSEs. A key challenge the regulation itself faces concerning said algorithms is ensuring that detection or labeling mechanisms for AI-generated synthetic media keep up with the pace of technological evolution in generative models (Yao et al., 2023). Similarly, legal ambiguity persists around accountability for automated content authorship as systems grow more autonomous (Federal Register, 2023; U.S. CopyrightOffice, 2023; European Parliament and Council, 2019; Leiser, 2023; Quintais, 2023; Keller, 2023a;b).

Risk assessments emphasize metaphorical 'virality' rather than actual generative and manipulation abilities, following Article 34 (European Parliament and Council, 2022a). Therefore, risks from realistic synthetic data escaping oversight and manipulation at scale could remain inadequately addressed. However, significant efforts have been made to enhance the European Commission's investigative powers in this regard. In March 2024, the Commission requested information and internal documents on risk assessments and mitigation measures for generative AI content.

Additionally, the Commission is working to identify electoral process integrity risks, issuing guidelines for VLOPs on mitigating election-related risks, including generative AI content (European Commission, 2024a). These guidelines represent best practices for managing electoral risks. If the Commission finds these measures inadequate, it can seek further information or initiate formal proceedings under the DSA. Despite these efforts, the challenge remains: how can regulatory action keep pace with the rapid development of harmful generative AI applications, and what tools can researchers use to ensure that the Commission's investigative and oversight powers are robust?

**Shortcomings Around Microtargeted Influence Campaigns.** While the DSA presents provisions for transparency in targeted advertising (Articles 15, 24, 42), the oversight enabling external auditing of aggregate messaging patterns and effects at population levels may still have limitations. There is a need for detailed guidance on how companies and assessors can implement risk assessments and auditing mechanisms. In particular, repositories would require more granular demographics beyond country locations, that would aid evaluating inequitable or unethical targeting.

Research from the AI Now Institute (2023) indicates that current algorithmic audits can reinforce private sector power, as companies often set the criteria and conduct the audits themselves. These audits tend to focus on technical evaluations of algorithms, neglecting broader socio-technical impacts like bias and discrimination. The lack of standardized definitions, methodologies, and benchmarks in the DSA's assessment process, at least since early adoption, complicates comparisons and reduces overall effectiveness. The presence of diverse services, features, and user bases of platforms further challenge the comparison of their impacts and harms. Moreover, the high threshold for reasonable assurance in audits, without clear standards, imposes significant risks on auditors and complicates the auditing of complex systems like recommendation algorithms.

When it comes to the Commission's regulatory efforts, there are frameworks supplementing existing legislation that ought to be mentioned in easing these limitations, such as the Delegated Act on Independent Audits under the DSA (European Commission, 2023a). These play an important role in providing additional guidance to providers of VLOPs and VLOSEs and auditing organisations in the preparation and issuance of audit reports and audit implementation reports, hence mitigating some of the shortcomings identified in the earliest versions of the DSA text. It remains to be established to what extent these guidelines are being followed in the relevant audits.

**Failures Addressing Filter Bubbles and Selective Exposure.** The DSA provides a necessary, yet insufficient, regulatory approach to issues related to filter bubbles, recommender distortions, or optimization risks that threaten exposure diversity – which have been linked to attitude polarization in scholarly research (Ross Arguedas et al., 2022) – despite associated calls (Liao & Fu, 2013). In particular, we observe how the DSA does not effectively cover the issues of content demotion and shadow banning, which often result in selective exposure of users to content and distorting the plurality of discourse online. While Articles 14 and 15 of the DSA mandate online platforms to clearly codify their content moderation rules, and Article 17 requires a 'Statement of Reasons' for every moderation action, critics have highlighted the lack of regulatory oversight for content demotion (Leerssen, 2023). Article 17 primarily focuses on notification, but it does not apply directly and explicitly to the problem of shadow-banning, which notoriously revolves around a lack of notification. Regulating demotion is technically complex due to the intricate nature of ranking algorithms and subsystems (Leerssen, 2023). Determining what constitutes demotion and assessing its impact presents significant challenges, requiring further research from the machine learning field. With the establishment of the AI Office, and the network of regulatory actors that coordinate the DSA's enforcement, greater technical capacity is required to fully understand and oversee platform ranking procedures (Novelli et al., 2024).

**Enabling Outrage-Focused Engagement Incentives.** In the DSA, interface dark patterns are only prohibited if directly 'deceptive' – an expansive standard under article 25 (European Parliament and Council, 2022a). Risk assessments, moreover, consider societal threats but not product choices that could psychologically manipulate users while technically remaining policy-compliant (following Article 34). An additional consideration is that beyond required periodic self-review on behalf of the platforms, which may be deemed an insufficient regulatory modality in itself, outrage-based engagement maximization risks remain unchecked. Therefore, the DSA might encounter some loopholes in its aim of comprehensively safeguard users or speech diversity against attention-captivating business models. Nevertheless, we observe promising interventions through the EU standard activities of CEN/CENELEC JTC 21 AI Group, now deriving technical work on ethical approaches to AI-enhanced nudging [WI=JT021006] (CEN/CENELEC, 2024).

## 4. Machine Learning Techniques for Assessing DSA Impacts

The DSA introduces new transparency and accountability obligations for online platforms, aiming to mitigate the spread of illegal content and protect users' fundamental rights. To effectively assess the real-world impacts of these provisions, researchers can leverage advanced machine learning techniques that enable large-scale analysis of complex socio-technical systems. In this section, we discuss the application of natural language processing, network modeling, and causal auditing methods to key research priorities for evaluating DSA implementation.

### 4.1. Evaluating Notice-and-Takedown Efficacy

One of the core pillars of the DSA is the requirement for platforms to establish clear and accessible mechanisms for users to report illegal content (Article 14). To assess the effectiveness of these notice-and-takedown systems, researchers

could employ natural language processing (NLP) techniques to analyze the vast amounts of data generated by platforms in compliance with DSA transparency requirements (Nenadic et al., 2023).

Advanced text classification models, such as transformer-based architectures, could be used to categorize user reports and platform responses by content type, language, and potential policy violations (European Commission, 2022). By comparing the volume and characteristics of notices before and after DSA implementation, researchers could identify changes in user reporting behavior and platform responsiveness attributable to the new legal framework (Trust Lab, 2023; Shao et al., 2019; Hsu et al., 2022). Moreover, sentiment analysis could be applied to assess the emotional valence of user reports and platform communications, providing insights into the perceived fairness and effectiveness of the notice-and-takedown process (Trust Lab, 2023; Humprecht et al., 2023). Qualitative methods, such as interviews and surveys with platform policy teams and content moderators, could complement these computational analyses by shedding light on the practical challenges and organizational dynamics shaping the implementation of DSA requirements (Altay et al., 2023).

### 4.2. Modeling disinformation Diffusion Patterns

Curbing the spread of online disinformation is a central objective of the DSA, requiring platforms to assess and mitigate systemic risks stemming from the design and operation of their services (Recital 9, Articles 34 and 35).

To support this goal, researchers can employ network modeling techniques to study the complex dynamics of information diffusion on social media platforms (Simpson & Yang, 2022). By representing users as nodes and their interactions (e.g., follows, shares, mentions) as edges, network models can simulate the spread of both legitimate and misleading content under different policy scenarios (Li & Chang, 2023; Jiang et al., 2023). These simulations can incorporate empirical data on user behavior and platform algorithms, as well as theoretical assumptions about the cognitive and social factors driving engagement with disinformation, such as confirmation bias and social identity (Kattenbeck & Elsweiler, 2019; Parvizi & Hmielowski, 2023; Blondé et al., 2022). Interrupted time series analysis can be used to compare the diffusion patterns of verified disinformation before and after the implementation of DSA-mandated risk mitigation measures, such as changes to recommender systems or the introduction of friction in sharing mechanisms (Bovet & Makse, 2018; Trust Lab, 2023). By identifying the most effective interventions for slowing the spread of false content, these models can inform evidence-based recommendations for platform design and content moderation policies (Luna, 2019).

### 4.3. Independently Auditing Advertising & Algorithms

The DSA seeks to enhance the transparency and accountability of online advertising systems, which have been criticized for enabling the microtargeting of harmful content and the amplification of extremist views (Hussein et al., 2020; Tomlein et al., 2021). To assess compliance with DSA requirements, such as the obligation to maintain advertising repositories (Article 39) and provide user-facing disclaimers, researchers can conduct independent audits using a combination of web scraping, crowdsourcing, and machine learning techniques.

For example, browser automation tools can be used to simulate user interactions with platform interfaces and collect data on the types of ads served to different demographic profiles (Laux et al., 2021). Computer vision algorithms can then be applied to classify the content and targeting parameters of these ads, enabling researchers to identify potential violations of DSA standards and discriminatory practices (Wilman, 2022). As an example, by comparing advertising patterns before and after DSA implementation, auditors could assess the effectiveness of the new transparency requirements in promoting user awareness and mitigating the risks of microtargeted manipulation.

Similar methods can be employed to audit the design and operation of algorithmic ranking and recommendation systems, which are subject to additional transparency obligations under the DSA (Article 39). By reverse-engineering the inputs and outputs of these systems through carefully designed experiments, researchers can infer the key factors driving content visibility and user engagement (Casper et al., 2024). These audits can help identify potential biases and filter bubble effects, informing the development of more diverse and inclusive algorithmic designs that align with DSA objectives. Another approach grounded in causal auditing is the use of counterfactual analysis, which could involve comparing the outcomes of different platform configurations or policy interventions while holding all other factors constant (Kluve et al., 2021). Researchers could use web scraping and browser automation tools to collect data on the ads served to different user profiles, and then use machine learning to generate counterfactual ad distributions that would have been shown under alternative targeting policies (Laux et al., 2021).

### 4.4. Enabling Broad Access to Transparency Data

To enhance transparency and allow for scrutiny of content moderation decisions, online platform providers must submit their statements of reasons to the DSA Transparency Database (European Commission, 2024b). This database enables almost real-time tracking of these decisions and offers various tools for accessing, analyzing, and downloading the information platforms provide when moderating con-

tent. Since its launch, the centralized database has attracted significant scholarly interest as an unprecedented source of data on real-world online moderation. However, research by Trujillo et al. (2024) indicates that large social media platforms have only partially adhered to the database's philosophy and structure. The authors highlight how, while all platforms met the DSA requirements, most omitted important yet optional details from their statements of reasons, thereby limiting the database's usefulness and requiring further analysis by researchers and policymakers.

In particular, to fully realize the potential of machine learning techniques for evaluating DSA impacts, researchers need access to comprehensive and representative data on platform operations and user behaviors. The DSA seeks to facilitate this by requiring platforms to provide vetted researchers with access to key metrics and datasets related to systemic risks and societal harms (Article 40) (Marrazzo, 2022). However, the practical implementation of these data sharing provisions could be hindered as platforms may seek to limit access based on commercial sensitivity or privacy concerns (Albert, 2023). To further overcome the identified barriers, researchers can advocate for the establishment of secure and privacy-preserving data sharing frameworks, such as the use of encrypted computation and synthetic data generation techniques (Marrazzo, 2022).

In parallel, researchers can leverage crowdsourcing and citizen science approaches to gather independent data on user experiences and platform practices (European Commission, 2022). By combining these bottom-up data collection efforts with the top-down transparency requirements of the DSA, researchers can build a more comprehensive and multifaceted evidence base for evaluating the regulation's impacts across different platforms, regions, and user groups.

## 5. Addressing Methodological Barriers

While the urgency of assessing the DSA using modern analytical methods remains clear given public stakes, meaningful challenges persist regarding ensuring thoughtful research design, obtaining representative data, and achieving generalizable insights across the sociotechnical diversity of platforms and national implementations.

### 5.1. Obtaining Adequate Research Access

Independent research hinges on appropriate data access unencumbered by proprietary gatekeeping. However, past voluntary transparency initiatives have often seen access increasingly restricted over time. Relying too optimistically on voluntary corporate cooperation, without the adequate policies and oversight, could threaten public interest audits. On top of that, there could be a lack of adequate provisions for specific research purposes under secure controls. Instituting data access mechanisms under democratic regulation, with compulsory sharing for specific research objectives, holds potential for significant impact. Yet, the revised Code of Practice on Disinformation has seen limited progress in this area (European Commission, 2022; Albert, 2023).

Independent scholarship around issues of advertising harms, algorithmic biases and disinformation dynamics hinges on appropriate data access for inference and auditing unencumbered by proprietary gatekeeping. Past failures of voluntary transparency initiatives highlight that absent assertive policy, dominant platforms tend to increasingly restrict third-party scrutiny through API limits or selective disclosures (Bruns, 2019; Marrazzo, 2022), jeopardizing capacities for public interest research. In this regard, non-commercial oversight bodies with mandated democratic governance may balance independence and access. Beyond basic provisions, the DSA should require extensive sharing of privacy-preserving synthetic datasets that maintain the original data distributions without revealing user information (Marrazzo, 2022). Combined with binding security commitments, these controlled data access agreements can facilitate external audits while preventing commercial exploitation or manipulation of algorithms

### 5.2. Isolating Effects of Multiple Emergent Regulations

The DSA entered into force alongside other regulations, such as the approved artificial intelligence (EU AI Act) and other texts regulating platform liability evolutions (DMA), data portability rights (Data Act), and algorithmic transparency requirements (GDPR), that shape key areas like advertising and content systems (European Parliament and Council, 2022b; European Parliament & Council, 2022; 2020; European Commission, 2016).

Isolating marginal effects specifically attributable to DSA provisions poses difficulties given overlapping phase-in timelines and pressures (Sullivan, 2023). Multi-pronged approaches can disentangle influences when evaluating observed platform changes. Interruptions time series modeling can estimate advertising transparency shifts before vs. after DSA rollout (Robertson et al., 2018). But since digital market regulations also compel disclosures (European Parliament and Council, 2022b), ethnographic observations within engineering teams would clarify relative weights of each intervention on infrastructure upgrades enabling oversight (Altay et al., 2023). Interpretative syntheses of textual regulatory histories provides contextual insights explaining the triggering of specific processes that quantitative signals may struggle to fully capture (Lejano, 2013). For example, the EU AI Act mandates algorithmic transparency provisions that likely reinforce pre-existing DSA auditing duties (European Parliament & Council, 2022), necessitating coordination to minimize double proceedings on potential

infractions.

### 5.3. Accounting for Regional Variations in DSA Implementation

One of the key challenges in evaluating the impact of the DSA is the regulation's decentralized enforcement structure, which relies on the cooperation and coordination of multiple actors at the EU and national levels. The DSA establishes a new European Board for Digital Services, composed of representatives from each Member State's Digital Services Coordinator (DSC), to oversee the consistent application of the regulation across the EU (Article 57). However, the primary responsibility for enforcing the DSA falls on the DSCs, who are appointed by their respective Member States and granted significant discretion in their supervisory and investigative powers (Articles 49-51).

This enforcement structure could have important implications for the regional variation in DSA implementation and enforcement. While the DSA sets out harmonized rules and obligations for platforms operating in the EU, the interpretation and application of these rules may differ depending on the resource availability, regulatory priorities, and institutional capacities of each Member State, as previously noted by legal scholars regarding this coordination for similar EU regulations such as the General Data Protection Regulation (GDPR) (Malgieri, 2019) and the AI Act (Novelli et al., 2024). For example, while DSCs in Western Europe may focus on mitigating foreign electoral interference, Eastern counterparts might prioritize domestic vaccine disinformation (Fourney et al., 2017; Forati & Ghose, 2021; Humprecht et al., 2023).

To capture these regional variations, researchers could employ comparative case study methods, analyzing the enforcement strategies and outcomes of different DSCs in relation to their national contexts (as for example done for associations laws and regimes in the EU member states (European Commission and Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs, 2022)). This can involve a combination of qualitative and quantitative approaches, such as interviews with DSC representatives, content analysis of enforcement decisions and public communications, and statistical modeling of complaint and takedown patterns across different Member States. For instance, researchers could compare the enforcement actions taken by the DSCs of Germany and France, two countries with divergent approaches to platform regulation. Germany has been a pioneer in this field, with its Network Enforcement Act (NetzDG) requiring platforms to remove illegal content within tight timeframes or face hefty fines (Net, 2017). In contrast, France has emphasized the role of self-regulation and collaborative governance, with its own law on online hate speech (Avia Law) being largely struck down by the Constitutional Council due to concerns over freedom of expression (Rights, 2020).

## 6. Call to Action for Coordinated Efforts

Realizing the DSA's immense potential for mitigating online harms requires actively fostering multi-stakeholder participation and coordinated mobilization of analytical capabilities spanning academia, civil society, platforms, journalists, policymakers and funding bodies.

### 6.1. Augmenting DSA Regulation with the EU AI Act

Our reflection on the DSA can extend to the current legislative ecosystem put forth by the EU. The EU AI Act, for example, proposes additional measures that, if implemented appropriately[2], may help mitigate certain limitations identified around generative models, targeting, filter bubbles, and engagement incentives:

- **Improving Oversight Over Generative Models and Synthetic Media.** The EU AI Act creates transparency obligations in Article 50, requiring certain AI systems interacting with people to indicate they are AI-generated. It also imposes record-keeping provisions in Article 12 for high-risk AI systems to facilitate traceability. Moreover, Article 52(4) empowers the Commission to designate general-purpose AI (GPAI) models as presenting systemic risks, even if they don't meet the initial threshold requirements. These articles could be complemented, through a relevant implementing act or further guidelines commissioned by the AI Office, to cover broader classes of synthetic media and generative models, particularly in light of the rapid development and potential impact of generative AI in social media contexts.

- **Guarding Against Microtargeted Influence Campaigns.** While the EU AI Act presently lacks binding assessments of model fairness or related provisions restricting unethical targeting, its Article 5 prohibits certain AI practices and allows expanding prohibited practices over time through delegated acts adopted by the Commission, as outlined in Article 7. If amended accordingly, outlawing clearly manipulative and discriminatory influence campaigns based on protected attributes may be more easily achieved.

[2]It should be noted that the EU AI Act adopts a risk-based approach, where certain provisions apply primarily to high-risk AI systems as delineated in the regulation. Yet AI systems not designated as high-risk may still enable concerning impacts, particularly in social media contexts (Hacker et al., 2023). Thus, when examining the potential to augment DSA regulation through the AI Act, it is relevant to contemplate measures appropriate not just for intrinsically high-risk systems, but also lower-risk classes enabling indirect societal harms.

- **Mitigating Filter Bubbles and Selective Exposure.** The EU AI Act's emphasis on risk management under Article 9 provides pathways for requiring impact assessments that could encompass threats to exposure diversity from recommender systems. Additional transparency obligations in Article 13 may also compel evaluating product choices endangering pluralistic discourse.

- **Limiting False Engagement Incentives.** While addictiveness protections are not explicitly mentioned in its text, as previously mentioned the EU AI Act enables revisiting permitted practices, as per Article 7, by updating the list in Annex III by adding high-risk AI systems. If outrage-based business models are finally shown to manipulate users or undermine speech diversity, and subsequently deemed to pose a risk to fundamental rights, prohibiting associated optimization targets may become appropriate. Beside that, the risk management obligations under Article 9 could be leveraged to require providers to assess and mitigate the risks associated with false engagement incentives, as these could be considered reasonably foreseeable misuses of high-risk AI systems.

### 6.2. Ensuring Multi-Stakeholder Participation

A robust DSA impact assessment necessitates inputs from diverse experts and communities representing the plurality of interests touched by platform governance decisions (Benkler et al., 2018). For example, survey instruments distributed to platform integrity teams, journalists, and end-users are all indispensable for holistically monitoring perceived shifts in disinformation prevalence, news media sustainability, and public awareness of manipulation risks before vs. after implementation (Papaevangelou, 2023; Cauffman & Goanta, 2021; Balkin, 2017). No single methodology in isolation can address multiply situated open questions, from quantifying policy compliance levels to explaining norm cascades in content moderation practices following heightened European scrutiny. Combining computational analytics of behavior changes with ethnographic observations within responding teams might aid in reconciling discrepancies between procedural outputs and on-the-ground challenges.

### 6.3. Coordination Across Disciplines

Assessments must also coordinate across academic disciplines, as legal scholars, political scientists, economists, psychologists and domain experts all offer indispensable analytical lenses into systemic impacts. For example, network analyses quantifying coordination between regulatory bodies requires pairing with interview data from policymakers on impediments to harmonization. And controlled behavioral experiments on disinformation warnings should inform technical design improvements of notification systems. Funding programs specifically encouraging cross-disciplinary investigation of platform governance impacts, potentially attentive to DSA-specific calls, can incentivize scientific multiplicity fit for complex sociotechnical responses (e.g., CiTIUS (2022)).

### 6.4. Grounding Policy in Evidence

Beyond enabling research, decision-making processes around issues like standardizing notice-and-takedown procedures, updating risk assessment methodologies, or revising independent audit protocols must continuously integrate emerging empirical insights in constructive, transparent ways that balance stakeholder interests and public safety priorities (Hacker et al., 2023; Sullivan, 2023). Adaptive governance would see European authorities and standardization bodies convene researchers and practitioners into regular working groups tasked with translating findings into actionable protocols and policy recommendations – a promising trajectory currently held by the heterogeneity of professional profiles within the CEN-CENELEC (2020).

## 7. Conclusion

As online platforms confront rising pressures over opaque methods and unchecked harms, policies mandating transparency and accountability warrant proactive, evidence-based evaluation balancing effectiveness and proportionality for continued innovation.

We argue that machine learning has significant untapped potential for quantifying the impacts of complex regulations like the EU Digital Services Act on various aspects, including content moderation, advertising, and recommendations. Detailed analysis revealed a gap between some overarching aims and granular research into on-ground impacts necessary for iterative refinement. We outline potential applications, from content analysis to causal auditing and diffusion modeling, to illuminate questions about the DSA's real-world impact. Coordinated efforts are needed to address methodological barriers related to appropriate data access, isolating marginal effects, and enabling generalization in order to facilitate credible assessment.

Overall, machine learning researchers should expand analytical techniques to catalyze evidence-based understanding of policy impacts on societal outcomes and online harms. Maintaining constructive partnerships between scientific and governing institutions is crucial as policymakers regulate previously unchecked technology sectors to serve the public interest.

## Impact Statement

This paper advocates using machine learning to evaluate online platform regulations, aiming to strengthen account-

ability around disinformation. We acknowledge responsible analytics requires safeguards against misuse and oversight balancing competing priorities. As platforms interacts with society at large, poorly designed policies could have significant unintended consequences.

Regulatory changes should be evidence-based, ethically vetted, and include affected communities. We explore this high-stakes issue to encourage the prudent and participatory use of ML in assessing complex sociotechnical policies for the public good.

## Acknowledgments

This work is supported by the EUHORIZON2021 European Union's Horizon Europe research and innovation programme (https://cordis.europa.eu/project/id/101073351/es) the Marie Skłodowska-Curie Grant No.: 101073351. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them. The authors have no relevant financial or non-financial interests to disclose.

## References

Gesetz zur verbesserung der rechtsdurchsetzung in sozialen netzwerken (netzwerkdurchsetzungsgesetz), 2017. URL https://www.bmj.de/SharedDocs/Downloads/DE/Gesetzgebung/RefE/NetzDG_engl.pdf?__blob=publicationFile&v=4. Unofficial translation. Last updated: 2017-07-12.

Acemoglu, D. and Ozdaglar, A. E. Opinion dynamics and learning in social networks. *Dyn. Games Appl.*, 1(1): 3–49, 2011. doi: 10.1007/S13235-010-0004-1. URL https://doi.org/10.1007/s13235-010-0004-1.

AI Now Institute. Algorithmic accountability: Moving beyond audits, 2023. URL https://ainowinstitute.org/publication/algorithmic-accountability.

Albert, J. Platforms' promises to researchers: First reports missing the baseline, Feb 2023. URL https://algorithmwatch.org/en/platforms-promises-to-researchers/.

Altay, S., Berriche, M., Heuer, H., Farkas, J., and Rathje, S. A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review*, 4(4):1–34, 2023. doi: 7910/DVN/5C3FHW.

Asnani, V., Yin, X., Hassner, T., and Liu, X. Reverse engineering of generative models: Inferring model hyperparameters from generated images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12):15477–15493, 2023. doi: 10.1109/TPAMI.2023.3301451. URL https://doi.org/10.1109/TPAMI.2023.3301451.

Bakir, V. Psychological operations in digital political campaigns: Assessing cambridge analytica's psychographic profiling and targeting. *Frontiers in Communication*, 5: 67, 2020.

Balkin, J. M. Free speech in the algorithmic society: Big data, private governance, and new school speech regulation. *UCDL rev.*, 51:1149, 2017. doi: 10.2139/ssrn.3038939.

Benkler, Y., Faris, R., and Roberts, H. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press, 11 2018. ISBN 9780190923624. doi: 10.1093/oso/9780190923624.001.0001. URL https://doi.org/10.1093/oso/9780190923624.001.0001.

Blondé, J., Easterbrook, M. J., Harris, P. R., Girandola, F., and Khalafian, A. Taking advantage of multiple identities to reduce defensiveness to personally threatening health messages. *Applied Psychology: Health and Well-Being*, 14(3):862–880, 2022. doi: 10.1111/aphw.12355.

Borgesius, F., Trilling, D., Moeller, J., Bodo, B., de Vreese, C., Helberger, N., and Review, I. Should we worry about filter bubbles? *Internet Policy Review*, Volume 5, 03 2016. doi: 10.14763/2016.1.401.

Bovet, A. and Makse, H. A. Influence of fake news in twitter during the 2016 US presidential election. *CoRR*, abs/1803.08491, 2018. URL http://arxiv.org/abs/1803.08491.

Bozdag, E. and van den Hoven, J. Breaking the filter bubble: democracy and design. *Ethics Inf. Technol.*, 17(4):249–265, 2015. doi: 10.1007/S10676-015-9380-Y. URL https://doi.org/10.1007/s10676-015-9380-y.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., McCandlish, . S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

Bruns, A. After the 'APIcalypse': social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22(11):1544–1566, 2019. doi: 10.1080/1369118X.2019.1637447. URL https://doi.org/10.1080/1369118X.2019.1637447.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Hagen, M. V., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., Krueger, D., and Hadfield-Menell, D. Black-box access is insufficient for rigorous ai audits, 2024.

Cauffman, C. and Goanta, C. A new order: The digital services act and consumer protection. *European Journal of Risk Regulation*, 12(4):758–774, 2021.

CEN-CENELEC. Focus group report - road map on artificial intelligence (ai), 2020. URL https://www.standict.eu/node/4854. Accessed on: 2024-04-30.

CEN/CENELEC. CEN/CLC/JTC 21 Artificial Intelligence technical work programme, 2024. URL https://standards.cencenelec.eu/dyn/www/f?p=205%3A22%3A0%3A%3A%3A%3AFSP_ORG_ID%2CFSP_LANG_ID%3A2916257%2C25&cs=1827B89DA69577BF3631EE2B6070F207D.

Choi, J. and Lim, K. Identifying machine learning techniques for classification of target advertising. *ICT Express*, 6(3):175–180, 2020. doi: 10.1016/J.ICTE.2020.04.012. URL https://doi.org/10.1016/j.icte.2020.04.012.

CiTIUS. Hybrid intelligence to monitor, promote and analyse transformations in good democracy practices. CORDIS - EU Research Projects, 2022. URL https://cordis.europa.eu/project/id/101073351. Grant agreement ID: 101073351, Reference Code: 101073351, Funded under: Horizon Europe, Marie Skłodowska-Curie Actions (MSCA), Doctoral Networks, European Union.

Da San Martino, G., Cresci, S., Barrón-Cedeño, A., Yu, S., Di Pietro, R., and Nakov, P. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. ISBN 9780999241165.

De Biasio, A., Monaro, M., Oneto, L., Ballan, L., and Navarin, N. On the problem of recommendation for sensitive users and influential items: Simultaneously maintaining interest and diversity. *Knowledge-Based Systems*, pp. 110699, 2023.

Du, M., Pentyala, S. K., Li, Y., and Hu, X. Towards generalizable deepfake detection with locality-aware autoencoder. In d'Aquin, M., Dietze, S., Hauff, C., Curry, E., and Cudré-Mauroux, P. (eds.), *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pp. 325–334. ACM, 2020. ISBN 978-1-4503-6859-9. doi: 10.1145/3340531.3411892. URL https://doi.org/10.1145/3340531.3411892.

European Commission. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj.

European Commission. The strengthened code of practice on disinformation 2022, 2022. URL https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation.

European Commission. Delegated regulation on independent audits under the digital services act, 2023a. URL https://digital-strategy.ec.europa.eu/en/library/delegated-regulation-independent-audits-under-digital-services-act.

European Commission. The digital services act package, 2023b. URL https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.

European Commission. European commission press release ip/24/1707, 2024a. URL https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1707.

European Commission. Dsa transparency database, 2024b. URL https://transparency.dsa.ec.europa.eu/. Accessed: 2024-06-05.

European Commission and Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs. *Comparative legal analysis of associations laws and regimes in the EU – Final report*. Publications Office of the European Union, 2022. doi: doi/10.2873/05056.

European Parliament and Council. Proposal for a regulation of the european parliament and of the council on european data governance (data governance act), 2020.

European Parliament and Council. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts - general approach, 2022.

European Parliament and Council. Directive (eu) 2019/790 of the european parliament and of the council of 17 april 2019 on copyright and related rights in the digital single market and amending directives 96/9/ec and 2001/29/ec, May 2019.

European Parliament and Council. Regulation (eu) 2022/2065 of the european parliament and of the council of 19 october 2022 on a single market for digital services and amending directive 2000/31/ec (digital services act). *Official Journal of the European Union*, L 367:1–98, 2022a. URL https://eur-lex.europa.eu/eli/reg/2022/2065/oj.

European Parliament and Council. Regulation (eu) 2022/1925 of the european parliament and of the council of 14 september 2022 on contestable and fair markets in the digital sector and amending directives (eu) 2019/1937 and (eu) 2020/1828 (digital markets act). Official Journal of the European Union, 2022b. URL https://eur-lex.europa.eu/eli/reg/2022/1925/oj. Text with EEA relevance.

European Parliament and Council. Regulation (eu) 2024/900 of the european parliament and of the council of 13 march 2024 on the transparency and targeting of political advertising. *Official Journal of the European Union*, L 900, 2024. URL https://eur-lex.europa.eu/eli/reg/2024/900/oj.

Federal Register. Copyright registration guidance: Works containing material generated by artificial intelligence. *U.S. Copyright Office, Library of Congress*, Mar 2023. URL https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence.

Forati, A. M. and Ghose, R. Geospatial analysis of misinformation in covid-19 related tweets. *Applied Geography*, 133:102473, 2021. ISSN 0143-6228. doi: https://doi.org/10.1016/j.apgeog.2021.102473. URL https://www.sciencedirect.com/science/article/pii/S0143622821000898.

Fourney, A., Rácz, M. Z., Ranade, G., Mobius, M., and Horvitz, E. Geographic and temporal trends in fake news consumption during the 2016 US presidential election. In Lim, E., Winslett, M., Sanderson, M., Fu, A. W., Sun, J., Culpepper, J. S., Lo, E., Ho, J. C., Donato, D., Agrawal, R., Zheng, Y., Castillo, C., Sun, A., Tseng, V. S., and Li, C. (eds.), *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pp. 2071–2074. ACM, 2017. ISBN 978-1-4503-4918-5. doi: 10.1145/3132847.3133147. URL https://doi.org/10.1145/3132847.3133147.

Gharibshah, Z., Zhu, X., Hainline, A., and Conway, M. Deep learning for user interest and response prediction in online display advertising. *Data Sci. Eng.*, 5(1):12–26, 2020. doi: 10.1007/S41019-019-00115-Y. URL https://doi.org/10.1007/s41019-019-00115-y.

Hacker, P., Engel, A., and Mauer, M. Regulating chatgpt and other large generative ai models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 1112–1123, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594067. URL https://doi.org/10.1145/3593013.3594067.

Haim, M., Graefe, A., and Brosius, H.-B. Burst of the filter bubble?: Effects of personalization on the diversity of google news. *Digital Journalism*, 6:330–343, 03 2018. doi: 10.1080/21670811.2017.1338145.

Heikkilä, M. We are hurtling toward a glitchy, spammy, scammy, ai-powered internet. *MIT Technology Review*, Apr 2023. URL https://www.technologyreview.com/2023/04/04/1070938/we-are-hurtling-toward-a-glitchy-spammy-scammy-ai-powered-internet/?

Helberger, N. On the democratic role of news recommenders. *Digital Journalism*, 7(8):993–1012, 2019. doi: 10.1080/21670811.2019.1623700. URL https://doi.org/10.1080/21670811.2019.1623700.

Heldt, A. P. *EU Digital Services Act: The White Hope of Intermediary Regulation*, pp. 69–84. Springer International Publishing, Cham, 2022. ISBN 978-3-030-95220-4. doi: 10.1007/978-3-030-95220-4_4. URL https://doi.org/10.1007/978-3-030-95220-4_4.

Hinds, J., Williams, E. J., and Joinson, A. N. "it wouldn't happen to me": Privacy concerns and perspectives following the cambridge analytica scandal. *Int. J. Hum. Comput. Stud.*, 143:102498, 2020. doi: 10.1016/J.IJHCS.2020.102498. URL https://doi.org/10.1016/j.ijhcs.2020.102498.

Hsu, C., Tsai, P., Yeh, T., and Hou, X. A comprehensive study of spatiotemporal feature learning for social medial popularity prediction. In Magalhães, J., Bimbo, A. D., Satoh, S., Sebe, N., Alameda-Pineda, X., Jin, Q., Oria, V., and Toni, L. (eds.), *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pp. 7130–7134. ACM, 2022. ISBN 978-1-4503-9203-7. doi: 10.1145/3503161.3551593. URL https://doi.org/10.1145/3503161.3551593.

Humprecht, E., Esser, F., Aelst, P. V., Staender, A., and Morosoli, S. The sharing of disinformation in cross-national comparison: analyzing patterns of resilience. *Information, Communication & Society*, 26(7):1342–1362, 2023. doi: 10.1080/1369118X.2021.2006744.

Hussein, E., Juneja, P., and Mitra, T. Measuring misinformation in video search platforms: An audit study on youtube. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–27, 2020.

Jaursch, J. Here is why digital services coordinators should establish strong research and data units - dsa observatory, Mar 2023. URL https://dsa-observatory.eu/2023/03/10/here-is-why-digital-services-coordinators-should-establish-strong-research-and-data-units/.

Jiang, C., Yu, Y., and Zhang, X. Modelling and analysis of misinformation diffusion based on the double intervention mechanism. *Journal of Information Science*, pp. 01655515231182076, 2023. doi: 10.1177/01655515231182076.

Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., and Bertulfo, D. J. The ghost in the machine has an american accent: value conflict in GPT-3. *CoRR*, abs/2203.07785, 2022. doi: 10.48550/ARXIV.2203.07785. URL https://doi.org/10.48550/arXiv.2203.07785.

Kattenbeck, M. and Elsweiler, D. Understanding credibility judgements for web search snippets. *Aslib Journal of Information Management*, 71(3):368–391, 2019. doi: 10.1108/AJIM-07-2018-0181.

Kaye, B. Australian mayor readies world's first defamation lawsuit over chatgpt. *Reuters*, Apr 2023. URL https://www.reuters.com/technology/australian-mayor-readies-worlds-first-defamation-lawsuit-over-chatgpt-content-2023-04-05/.

Keller, P. Generative ai and copyright: Convergence of opt-outs? *Kluwer Copyright Blog*, Nov 2023a. URL https://copyrightblog.kluweriplaw.com/2023/11/23/generative-ai-and-copyright-convergence-of-opt-outs/.

Keller, P. A first look at the copyright relevant parts in the final ai act compromise. *Kluwer Copyright Blog*, Dec 2023b. URL https://copyrightblog.kluweriplaw.com/2023/12/11/a-first-look-at-the-copyright-relevant-parts-in-the-final-ai-act-compromise/#_ftnref1.

Kelly, S. M. Snapchat's new ai chatbot is already raising alarms among teens. *CNN*, Apr 2023. URL https://edition.cnn.com/2023/04/27/tech/snapchat-my-ai-concerns-wellness/index.html.

Khoo, B. B. G., Phan, R. C., and Lim, C. H. Deepfake attribution: On the source identification of artificially generated images. *WIREs Data Mining Knowl. Discov.*, 12(3), 2022. doi: 10.1002/WIDM.1438. URL https://doi.org/10.1002/widm.1438.

Kirchner, J. and Reuter, C. Countering fake news: A comparison of possible solutions regarding user acceptance and effectiveness. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW2):1–27, 2020.

Klonick, K. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.*, 131:1598, 2017. URL https://harvardlawreview.org/print/vol-131/the-new-governors-the-people-rules-and-processes-governing-online-speech/.

Kluve, J., Naldini, A., and Pompili, M. *Design and commissioning of counterfactual impact evaluations: A practical guidance for ESF managing authorities*. Publications Office of the European Union, 2021. ISBN 9789276407256. doi: 10.2767/02762. URL https://op.europa.eu/en/publication-detail/-/publication/dd4a4fc7-42a3-11ec-89db-01aa75ed71a1.

Kuczerawy, A. The good samaritan that wasn't: voluntary monitoring under the (draft) digital services act, 2021.

Laux, J., Wachter, S., and Mittelstadt, B. Taming the few: Platform regulation, independent audits, and the risks of capture created by the DMA and DSA. *Computer Law & Security Review*, 43:105613, 2021. ISSN 0267-3649. doi: https://doi.org/10.1016/j.clsr.2021.105613. URL https://www.sciencedirect.com/science/article/pii/S0267364921000868.

Leerssen, P. An end to shadow banning? transparency rights in the digital services act between content moderation and curation. *Computer Law & Security Review*, 48:105790, 2023. ISSN 0267-3649. doi: https://doi.org/10.1016/j.clsr.2023.105790. URL https://www.sciencedirect.com/science/article/pii/S0267364923000018.

Leiser, M. N. Reimagining Digital Governance: The EU's Digital Service Act and the Fight Against Disinformation. *SSRN*, 2023. doi: https://dx.doi.org/10.2139/ssrn.4427493.

Lejano, R. *Frameworks for policy analysis: Merging text and context*. Routledge, 2013. doi: 10.4324/9780203625422.

Li, J. and Chang, X. Combating misinformation by sharing the truth: a study on the spread of fact-checks on social media. *Information systems frontiers*, 25(4):1479–1493, 2023.

Liao, Q. V. and Fu, W.-T. Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 2359–2368, 2013.

Lo, V.-H. and Wei, R. Modeling the dynamic process and adverse effects of misinformation. In *Miscommunicating the COVID-19 Pandemic*, pp. 196–207. Routledge, 2023.

Lopatto, E. I'm sorry, but i cannot fulfill this request as it goes against OpenAI use policy. *The Verge*, Jan 2024. URL https://www.theverge.com/2024/1/12/24036156/openai-policy-amazon-ai-listings.

Luna, F. Identifying and evaluating layers of vulnerability–a way forward. *developing world bioethics*, 19(2):86–95, 2019. doi: 10.1111/dewb.12206.

Malgieri, G. Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations. *Computer Law & Security Review*, 35(5):105327, October 2019. ISSN 0267-3649. doi: 10.1016/j.clsr.2019.05.002. URL https://www.sciencedirect.com/science/article/pii/S0267364918303753.

Marrazzo, F. Doing research with online platforms: An emerging issue network. In *Handbook of research on advanced research methodologies for a digital society*, pp. 65–86. IGI Global, 2022. doi: 10.4018/978-1-7998-8473-6.ch006.

Matz, S. C. Personal echo chambers: Openness-to-experience is linked to higher levels of psychological interest diversity in large-scale behavioral data. *Journal of Personality and Social Psychology*, 121(6):1284, 2021. doi: 10.1037/pspp0000324.

Mavriki, P. and Karyda, M. Big data analytics: From threatening privacy to challenging democracy. In *E-Democracy–Safeguarding Democracy and Human Rights in the Digital Age: 8th International Conference, e-Democracy 2019, Athens, Greece, December 12-13, 2019, Proceedings 8*, pp. 3–17. Springer, 2020.

Mirsky, Y. and Lee, W. The creation and detection of deepfakes: A survey. *ACM Comput. Surv.*, 54(1):7:1–7:41, 2022. doi: 10.1145/3425780. URL https://doi.org/10.1145/3425780.

MistralAI. Bringing open ai models to the frontier. *Mistral AI — Open-weight models*, Jan 2024. URL https://mistral.ai/news/about-mistral-ai/.

Mok, A. This ai stock trader engaged in insider trading - despite being instructed not to – and lied about it. *Business Insider*, Dec 2023. URL https://www.businessinsider.com/ai-deceive-users-insider-trading-study-gpt-2023-12.

Morrison, S. How unbelievably realistic fake images could take over the internet. *Vox*, Mar 2023. URL https://www.vox.com/technology/2023/3/30/23662292/ai-image-dalle-openai-midjourney-pope-jacket.

Nannini, L. Voluminous yet vacuous? semantic capital in an age of large language models. In Ganapini, M. B., Loreggia, A., Mattei, N., Rossi, F., Srivastava, B., and Venable, K. B. (eds.), *Proceedings of the Workshop on Ethics and Trust in Human-AI Collaboration: Socio-Technical Approaches (ETHAICS 2023) co-located with 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023) Macao, August 21, 2023., Macao, August 21, 2023*, volume 3547 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2023. URL https://ceur-ws.org/Vol-3547/paper5.pdf.

Nannini, L., Bonel, E., Bassi, D., and Maggini, M. J. Beyond phase-in: assessing impacts on disinformation of the eu digital services act. *AI and Ethics*, 2024. ISSN 2730-5961. doi: 10.1007/s43681-024-00467-w. URL https://doi.org/10.1007/s43681-024-00467-w.

Neff, G. The new digital dark age. *Wired*, Jan 2024. URL https://www.wired.com/story/the-new-digital-dark-age/#intcid=_wired-bottom-recirc-v2_7e24dea7-af8f-4bb7-b40c-ea60b00a0ec5_cral2-2-reranked-by-vidi_fallback_popular4-1.

Nenadic, I., Brogi, E., and BLEYER-SIMON, K. Structural indicators to assess effectiveness of the eu's code of practice on disinformation. Technical report, Centre for Media Pluralism and Media Freedom, [European Digital Media Observatory (EDMO)], 2023.

Notopoulos, K. A car dealership added an ai chatbot to its site. then all hell broke loose. *Business Insider*, Dec 2023. URL https://www.businessinsider.com/car-dealership-chevrolet-chatbot-chatgpt-pranks-chevy-2023-12.

Novelli, C., Hacker, P., Morley, J., Trondal, J., and Floridi, L. A robust governance for the ai act: Ai office, ai board, scientific panel, and national authorities. *Available at SSRN*, 2024. URL https://ssrn.com/abstract=4817755.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/arXiv.2303.08774.

Papaevangelou, C. The role of citizens in platform governance: A case study on public consultations regarding online content regulation in the european union. *Global Media and China*, 8(1):39–56, 2023. doi: 10.1177/20594364221150142.

Parvizi, J. and Hmielowski, J. D. Breaking the mold: examining the effectiveness of techniques to reduce motivated reasoning. *Atlantic Journal of Communication*, pp. 1–14, 2023. doi: 10.1080/15456870.2023.2224482.

Perrigo, B. Bing's ai is threatening users. that's no laughing matter. *Time*, Feb 2023. URL https://time.com/6256529/bing-openai-chatgpt-danger-alignment/.

Pham, A., Rubel, A., and Castro, C. Social media, emergent manipulation, and political legitimacy. In *The Philosophy of Online Manipulation*, pp. 353–369. Routledge, 2022. doi: 10.4324/9781003205425.

Pérez-Escoda, A., Boulos, S., Establés, M.-J., and García-Carretero, L. Polarization in media discourses on europeanization in spain. *Politics and Governance*, 11 (2):221–234, 2023. ISSN 2183-2463. doi: 10.17645/pag.v11i2.6419. URL https://www.cogitatiopress.com/politicsandgovernance/article/view/6419.

Quin, F., Weyns, D., Galster, M., and Silva, C. C. A/b testing: A systematic literature review. *Journal of Systems and Software*, 211:112011, 2024. ISSN 0164-1212. doi: https://doi.org/10.1016/j.jss.2024.112011. URL https://www.sciencedirect.com/science/article/pii/S0164121224000542.

Quintais, J. P. Generative ai, copyright and the ai act. *Kluwer Copyright Blog*, May 2023. URL https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/.

Rights, E. D. French avia law declared unconstitutional: what does this teach us at eu level?, 2020.

URL https://edri.org/our-work/french-avia-law-declared-unconstitutional-what-does-this-teach-us-at-eu-level/.

Robertson, R. E., Lazer, D., and Wilson, C. Auditing the personalization and composition of politically-related search engine results pages. In *Proceedings of the 2018 World Wide Web Conference*, pp. 955–965, 2018.

Romano, S., Kerby, N., Angius, R., Robutti, S., Schueler, M., Faddoul, M., Çetin, R. B., Helming, C., Müller, A., Spielkamp, M., and et al. Generative ai and elections: Are chatbots a reliable source of information for voters? *AIForensics, AlgorithmWatch*, Dec 2023. URL https://aiforensics.org/uploads/AIF_AW_Bing_Chat_Elections_Report_ca7200fe8d.pdf.

Ross Arguedas, A., Robertson, C., Fletcher, R., and Nielsen, R. Echo chambers, filter bubbles, and polarisation: A literature review. Technical report, Reuters Institute, 2022. URL https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review.

Sankaran, V. Chatgpt cooks up fake sexual harassment scandal, names real law professor as accused. *The Independent*, Apr 2023. URL https://www.independent.co.uk/tech/chatgpt-sexual-harassment-law-professor-b2315160.html.

Schäfer, S. Incidental news exposure in a digital media environment: a scoping review of recent research. *Annals of the International Communication Association*, 47(2): 242–260, 2023. doi: 10.1080/23808985.2023.2169953. URL https://doi.org/10.1080/23808985.2023.2169953.

Shao, J., Shen, H., Cao, Q., and Cheng, X. Temporal convolutional networks for popularity prediction of messages on social medias. In Zhang, Q., Liao, X., and Ren, Z. (eds.), *Information Retrieval - 25th China Conference, CCIR 2019, Fuzhou, China, September 20-22, 2019, Proceedings*, volume 11772 of *Lecture Notes in Computer Science*, pp. 135–147. Springer, 2019. ISBN 978-3-030-31623-5. doi: 10.1007/978-3-030-31624-2_11. URL https://doi.org/10.1007/978-3-030-31624-2_11.

Sharma, N., Liao, Q. V., and Xiao, Z. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642459. URL https://doi.org/10.1145/3613904.3642459.

Simpson, T. and Yang, F.-J. Some hands-on approaches to fake political news detection. In *Proceedings of the 2022 5th International Conference on Signal Processing and Machine Learning*, pp. 179–188, 2022.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In Guo, Y. and Farooq, F. (eds.), *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pp. 2239–2248. ACM, 2018. doi: 10.1145/3219819.3220046. URL https://doi.org/10.1145/3219819.3220046.

Spisak, J. and Edunov, S. The llama ecosystem: Past, present, and future. *AI at Meta*, Sep 2023. URL https://ai.meta.com/blog/llama-2-updates-connect-2023/.

Sullivan, D. Unpacking "systemic risk" under the eu's digital service act, Jul 2023. URL https://techpolicy.press/unpacking-systemic-risk-under-the-eus-digital-service-act/.

Tian, Y. and Wang, L. Dynamics of opinion formation, social power evolution, and naïve learning in social networks. *Annu. Rev. Control.*, 55:182–193, 2023. doi: 10.1016/J.ARCONTROL.2023.04.001. URL https://doi.org/10.1016/j.arcontrol.2023.04.001.

Tomlein, M., Pecher, B., Simko, J., Srba, I., Moro, R., Stefancova, E., Kompan, M., Hrckova, A., Podrouzek, J., and Bielikova, M. An audit of misinformation filter bubbles on youtube: Bubble bursting and recent behavior changes. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pp. 1–11, 2021. doi: https://doi.org/10.1145/3460231.3474241. URL https://arxiv.org/abs/2203.13769.

Trujillo, C., Sanchez, L., and Gomez, M. Evaluating the effectiveness of the dsa transparency database in content moderation: Challenges and opportunities, 2024. URL https://arxiv.org/abs/2312.10269v1.

Trust Lab. Code of practice on disinformation. a comparative analysis of the prevalence and sources of disinformation across major social media platforms in poland, slovakia, and spain. Technical report, European Union, 2023. URL https://disinfocode.eu/wp-content/uploads/2023/09/code-of-practice-on-disinformation-september-22-2023.pdf.

U.S. CopyrightOffice. U.s. copyright office fair use index. *U.S. Copyright Office*, Nov 2023. URL https://www.copyright.gov/fair-use/.

Wilman, F. The digital services act (dsa)-an ooverview. *Available at SSRN 4304586*, 2022. URL http://dx.doi.org/10.2139/ssrn.4304586.

Wolters, P. and Gellert, R. Towards a better notice and action mechanism in the dsa. *J. Intell. Prop. Info. Tech. & Elec. Com. L.*, 14:403, 2023.

Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y., and Chen, Z. How much can behavioral targeting help online advertising? In Quemada, J., León, G., Maarek, Y. S., and Nejdl, W. (eds.), *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pp. 261–270. ACM, 2009. ISBN 978-1-60558-487-4. doi: 10.1145/1526709.1526745. URL https://doi.org/10.1145/1526709.1526745.

Yao, H., Lou, J., Ren, K., and Qin, Z. Promptcare: Prompt copyright protection by watermark injection and verification, 2023.

Yu, N., Skripniuk, V., Abdelnabi, S., and Fritz, M. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 14428–14437. IEEE, 2021. ISBN 978-1-6654-2812-5. doi: 10.1109/ICCV48922.2021.01418. URL https://doi.org/10.1109/ICCV48922.2021.01418.

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., and Choudhury, M. D. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In Schmidt, A., Väänänen, K., Goyal, T., Kristensson, P. O., Peters, A., Mueller, S., Williamson, J. R., and Wilson, M. L. (eds.), *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pp. 436:1–436:20. ACM, 2023. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3581318. URL https://doi.org/10.1145/3544548.3581318.