AI and Ethics (2025) 5:1241–1269
https://doi.org/10.1007/s43681-024-00467-w

ORIGINAL RESEARCH

Beyond phase-in: assessing impacts on disinformation of the EU Digital Services Act

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

Received: 7 November 2023 / Accepted: 7 March 2024 / Published online: 11 April 2024
© The Author(s) 2024

Abstract
This work proposes a comprehensive research agenda to empirically evaluate the real-world impacts of the European Union's Digital Services Act (DSA) on combating online disinformation. It provides background on the DSA's context, mechanisms, timeline, and expected effects on platforms to situate the need for rigorous impact assessment. A detailed legal, technical, psychological, behavioral and ethical critique reveals meaningful gaps in the DSA requiring ongoing regulatory refinement and oversight. Most critically, the paper puts forth an encompassing framework spanning computational analytics, interviews, ethnography, surveys, discourse analysis and mixed methods to rigorously assess the DSA's multi-dimensional effects on complex factors enabling disinformation proliferation. Priorities include evaluating notice-and-takedown efficacy, advertising transparency improvements, risk assessment outcomes, oversight integration, and procedural shifts in platform governance. Coordinated efforts between researchers, regulators and platforms are needed to address methodological challenges around isolating DSA impacts amidst an evolving EU regulatory landscape, constrained data access from platforms, and difficulties generalizing findings across the sociotechnical diversity of platforms and national contexts in EU Member States.

Keywords Digital Services Act · Online platforms · Content moderation · Disinformation · Public policy evaluation

1 Introduction

With the European Union's Digital Services Act (referred to as the 'DSA') [57] entering into force in late 2022 and fully applicable from February 2024, a timely and urgent opportunity emerges to empirically evaluate if its mechanisms can meaningfully constrain opaque algorithms, advertisers and platforms implicated in online harms.

In particular, the rise of highly capable and scarcely regulated Artificial Intelligence (AI) applications on online platform raise urgent questions around how algorithmic governance must adapt to grapple with emerging societal risks [26, 81, 126, 127, 129, 150, 164, 179]. Particularly, recent advances in generative AI, driven by massive internet-trained language models, enable automated generation of text, images and other content that appears authentic [7, 16, 20, 32, 39, 73, 113, 132, 146, 171, 177, 188]. Nevertheless, these opaque systems also have the potential to generate manipulative content, disseminate harmful disinformation, and perpetuate systematic biases on a massive scale without meaningful human oversight or transparency [1, 13, 25, 40, 72, 80, 84, 95, 103, 117, 128, 131, 133, 151, 155, 156]. Precisely with this purpose of tackling issues such as disinformation and preventing harmful activities online, the DSA aims to ensure transparency across platforms, establish accountability mechanisms, and create specific provisions addressing emerging social risks posed by opaque algorithmic systems. For instance, it aims to address disinformation spread via algorithmic recommender systems, behavioral advertising, and electoral integrity.

✉ Luca Nannini
l.nannini@usc.es; lnannini@minsait.com

Eleonora Bonel
eleonora.bonel@sciencespo.fr

Davide Bassi
davide.bassi@usc.es

Michele Joshua Maggini
michelejoshua.maggini@usc.es

[1] Minsait, Indra Sistemas S.A., Madrid, Spain

[2] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain

[3] Institute of Political Studies, École d'affaires publiques, SciencesPo, Paris, France

However, comprehensive empirical assessments of the DSA's effectiveness in mitigating emerging social harms from AI and other widely deployed opaque algorithmic applications, remain lacking to date. Academic discourse has thus far only provided piecemeal or theoretical perspectives on the DSA's potential [61, 78, 82, 104, 108, 109, 169], rather than rigorously quantifying its multidimensional effects on complex factors enabling the phenomenon of disinformation. Potentially, this could be because it was premature to position the DSA's direct contribution to curbing online mis/disinformation.[1]

However, without comprehensive, empirical evaluation, the DSA's suitability for effective AI governance remains speculative. This is why the present research argues for the necessity of gathering robust evidence on the efficacy of DSA mechanisms through a range of quantitative and qualitative methods. This translates to the overarching research question: *How can comprehensive, empirical research effectively quantify and evaluate the real-world impacts of the DSA's mechanisms on combating online disinformation?*

Bridging this gap requires a three-pronged approach. First, providing extensive definitions of the DSA's key provisions, obligations, and timelines to situate these mechanisms. Second, articulating known critiques around the efficacy of the DSA's approach to curbing disinformation and broader policy aims. Third and centrally, proposing research approaches to shed lights over the DSA's impacts.

The methods proposed for evaluating DSA impact on online disinformation can be directly extended to examine its role in shaping policy responses to emerging AI challenges. For instance, the DSA's provisions around algorithmic transparency and behavioral targeting disclosure address emerging social risks from AI "black box" systems. Similarly, its risk assessment and risk mitigation requirements for "very large online platforms" (VLOPs) could mitigate potential harms from highly scalable applications of generative AI if extended beyond social media.

Far from thus providing definitive solutions, the intended contribution of this work stands in outlining a critical research agenda to inform incremental optimizations of the DSA based on empirically quantified, multidimensional insights into its provisions' efficacy.

The rest of the paper is structured as follows. In Sect. 2, we provide an Overview of the Digital Services Act and its categorization of online entities. Within, Sect. 2.2 elucidates the provisions on AI accountability. Critiques of the DSA's approach towards disinformation are articulated in Sect. 3, encompassing an analysis of the DSA's regulatory antecedents Sect. 3.1 and leading to a review of the DSA's persistent gaps in Sect. 3.2. The latter section encompasses legal, technical, psychological, behavioral, and ethical critiques. Section 4 outlines the Research Approaches to Evaluate DSA Impact on Disinformation, which further bifurcates into quantitative research directions in Sect. 4.1.1, qualitative research approaches in Sect. 4.1.2, and mixed methods research alongside methodological challenges in Sect. 4.1.3; research limitations are discussed in Sect. 4.2 before the Conclusion in Sect. 5.

2 The EU Digital Services Act framework and aims

The European Parliament passed the DSA in 2022 as a comprehensive legal framework to regulate digital platforms and services. This section provides first an overview of the key mechanisms (Sect. 2.1), then situating them with AI governance concerns within disinformation across online platforms (Sect. 2.2).

2.1 Overview of key mechanisms

Before analyzing intersections with AI governance and disinformation, it is useful to overview the DSA's key provisions and mechanisms instituting heightened accountability for online platforms. The interested reader will find a detailed analysis including its context and mechanisms within Appendix 2.

At its core, the DSA establishes transparency duties, risk analysis requirements and independent auditing processes for Very Large Online Platforms (VLOPs) and search engines reaching over 10% of the EU population—around 45 million monthly active users (see Recital 76). VLOPs like Google, Meta, X, and TikTok must implement strict transparency rules around ad targeting, content moderation, and data sharing with vetted researchers. They face recurring systemic risk assessments evaluating dangers to

[1] The present paper defines disinformation according to the European Democracy Action Plan's definition (EDAP) as "*false or misleading content that is spread with an intention to deceive or secure economic or political gain and which may cause public harm.*" [48]. While disinformation is often used interchangeably with misinformation in the literature, the latter presents a slightly different nuance. According to the EDAP, "*misinformation is false or misleading content shared without harmful intent though the effects can be still harmful, e.g. when people share false information with friends and family in good faith*". The nuance between the two terms lies in the intention for dissemination, which may be challenging to determine a priori. Establishing a legal definition of disinformation also posed a challenge in the context of the Digital Services Act (DSA), as policymakers aimed to ensure the term was comprehensive enough to distinguish it without over-regulating the information ecosystem [168]. Therefore, we acknowledge some of the challenges in strictly defining disinformation, as reflected in its different definitions in the EU regulatory landscape. However, for the purposes of this paper, we highlight that the various definitions of disinformation allow it to be characterized as an umbrella term. The interested reader can find a complete overview of the debate in Appendix 1.

Table 1 Mechanism codes for key Digital Services Act provisions

Code	Mechanism	Description	Articles
NTD	Notice-and-Takedown	Procedures for content removal requests, statement of reason and illegal content takedowns	Art. 14-15
AT	Advertising Transparency	Requirements for transparency about targeted online ads and the designation of independent Digital Services Coordinators	Art. 24, 30, 39
RA	Risk Assessments	Designation of VLOPs/VLOSEs and subsequent obligations for systemic risk assessments about content and algorithms	Art. 26-27, 33, 34-35
DD	Due Diligence	Graduated due diligence duties for online intermediaries and relevant provisions	Art. 10-15, 35-37, 44-48
AD	Auditing	Independent recurring audits evaluating compliance	Art. 28, 33, 37

electoral integrity, public discourse, user rights and equality emerging from their platforms' designs, business models and algorithms. Preventative and proportional risk mitigation duties follow uncovered threats. The DSA also compels VLOPs to undergo annual independent audits by accredited entities assessing procedural compliance in areas like content policies, advertising, algorithms and risk management. Audit summaries with key findings must become publicly accessible. Additionally, heightened due diligence obligations require VLOPs to bolster identity verification, external vetting of policies, and coordinated oversight across EU member states through a European Board for Digital Services [52]. Graduated responsibilities extend to smaller platforms as well, including accessible reporting procedures for illegal content.

To enable explicit mapping between key DSA mechanisms and subsequent critiques and research directions, core provisions are coded by unique abbreviations as summarized in Table 1. These mechanism codes will be referenced throughout the text where relevant.

2.2 DSA interventions for AI governance

While not positioning itself as an "AI regulator" outright—as expected by the EU's AI Act [37]—specific DSA transparency provisions granting external data access, requiring recurring independent audits evaluating algorithmic systems, mandating systemic risk assessments, and restricting manipulative "dark patterns" (Art.25) align closely with emerging policy needs around accountable and ethical AI.

Rather than examining individual AI applications in isolation, it is more constructive to review promising pathways the DSA establishes to institute sustained external oversight and accountability over influential algorithmic systems across sectors. Four key mechanisms warrant consideration beyond current online platforms:

Qualified Data Access Article 31 of the DSA mandates VLOPs to grant access to datasets and platform data to vetted researchers investigating manifest or foreseeable systemic risks from their services, subject to confidentiality safeguards, as it recognizes that "*access to data is needed to fully understand the systemic risks posed by online platforms*" (Recital 58). This enables independent analysis in the public interest examining processes and patterns underlying concerns like optimizing algorithms disproportionately amplifying inflammatory content (Code: AD). A similar transparency approach proactively requiring developers of influential AI systems including generative models to provide external researchers carefully controlled and purpose-limited data access would be invaluable. As Recital 61 states, "*access to data is needed for researching, understanding and addressing societal issues like algorithmic amplification*" hence provisions enabling examining aspects like biases encoded within foundation datasets scraped non-consensually from the web are important. For instance, researchers could probe issues like demographic representativeness and known societal biases using testing datasets (Article 31).

Such qualified external data access provisions significantly enhance accountability over AI risks that remain inscrutable within opaque systems (Recital 57), ensuring independent research in the public interest is not obstructed by default. As Recital 58 states, it balances openness to scrutiny with reasonable confidentiality safeguards, facilitating identification of model deficiencies before broad proliferation. Furthermore, as Recital 70d of the EU AI Act states [37], provisions enabling scrutiny of opaque AI systems align with DSA risks assessments detecting manipulated content—hence qualified access for examining issues like generative models encoding biases facilitates policy coherence

Algorithmic Audits The annual independent DSA audits analyzing VLOPs' algorithmic systems including content curation, recommender processes and advertising auctions (Art.37) offer valuable templates for instituting sustained external oversight assessing influential AI across sectors. Tailored auditing protocols and benchmarks evaluating aspects like model robustness, representational bias, classification skew, data security, safeguards against misuse, and model limitations mandated through governance measures would be transformative (Code: AD). The AI Act (Recital 60v) [37] particularly underscores DSA's transparency duties on platforms to mitigate risks from disseminating

synthetically generated content—thereby sustained external audits provide oversight regarding risks like generative models propagating information disorders if deployed at-scale.

For instance, the exponential rise of services employing Generative AI e.g. OpenAI's ChatGPT,[2] potentially exacerbating threats around disinformation and confidential data exposure [62] underscores the urgency of instituting recurring external oversight before unforeseen harms emerge at population scale.

Risk Assessments The required recurring DSA risk assessments analyzing threats to electoral processes and vulnerable groups enabled by VLOP systems including proprietary algorithms (Art. 34, 35) highlight the necessity of mandating that AI developers comprehensively document potential dangers from creating addictive, biased or manipulative systems. As generative models like text and media synthesizers scale exponentially in sophistication, developers should increasingly bear the onus of demonstrating diligent inquiries evaluating capacities for harm. Indeed, relying purely on voluntary self-assessment of risks, in the context of the unprecedented acceleration in generative AI capabilities presaging looming threats from fake accounts, spam, phishing campaigns, disinformation floods, would prove inadequate. In this vein, the AI Act (especially Recitals 60v, 70d) [37] complements the DSA's transparency obligations on Platforms to mitigate risks from disseminating synthetically generated content, as both texts call out the specific need to adequately audit algorithms. Moreover, the AI Act's recitals and DSA provisions would seem to highlight how sustained external audits provide oversight regarding risks like generative models propagating information disorders if deployed at-scale. Therefore, reporting mechanism for users could be imposed for detecting harmful, biased or abusive outputs from AI systems like generative models deployed on platforms (Code: NTD). Systematizing evaluation of abuse potentials through mandatory issue scoping anchored to sectoral impact assessments tailored for AI systems would constitute measurable progress. Thereby the DSA's general approach requiring recurrent evaluations of risk trajectories (Code: RA) provides a broadly extensible template for AI governance.

[2] The popularity of ChatGPT represents a prime case revealing policy gaps regarding influential AI systems [158, 159, 166]. Harm considerations include exposure of personal information through conversations; reinforcement of biases from web-scraped training data; overstated confidence in responses leading to public disinformation; potential to automate phishing, spam, misinformation campaigns; exacerbating cheating in academia and assessments; facilitating creation of malware/ransomware; legal risks regarding IP, defamation and more [78, 95, 133, 142, 156, 165, 179, 186]. However, ChatGPT currently operate outside regulatory purview of measures like the DSA tailored to dominant platforms, as observed by [78].

Data Sharing Duties Article 31 introduces duties around information sharing and coordinated incident responses, mandating VLOPs to notify authorities regarding manifest illegal content risks, to provide additional data about ad sponsors, and to report security compromises enabling systemic harms (Code: DD, AT). In the context of harmful AI applications, establishing clear obligations around risk communication and coordinating disclosures following harm incidents would aid a collective response. For instance, requirements around expeditiously notifying relevant authorities if deployed generative chatbots demonstrably enable confidential data leaks (Recital 58) can enable interventions limiting damage, while avoiding knee-jerk reactions. As Recital 61 states, pragmatic coordination mechanisms recognizing anticipating challenges from technologies operating at immense scale would underscore shared responsibilities in averting crises, with the relevant notices being submitted by trusted flaggers. Following the DSA's approach of instituting oversight and communication channels by design (Article 31) (Code: DD) extends naturally to population-scale AI systems.

3 Critiques of the DSA's approach to disinformation

While the Digital Services Act establishes more extensive obligations on platforms compared to prior self-regulatory efforts, open questions remain regarding its efficacy and limitations. Having outlined the DSA's core provisions as well as touchpoints to AI governance, this section analyzes weaknesses in previous self-governance efforts. Based on these prior gaps, the section then provides an in-depth critique of the DSA across interrelated dimensions—legal, psychological, and ethical—to reveal inconsistencies and unexamined assumptions that warrant ongoing empirical assessment and regulatory refinement.

Specifically, the trajectory leading up to the DSA represents a gradual acknowledgment of self-governance limitations around complex threats like disinformation, giving way to formal multi-stakeholder cooperation and binding regulations. However, tensions persist around reconciling such binding rules with evidentiary burdens, accuracy, oversight, and proportionality amidst pressing societal calls for action.

3.1 Regulatory antecedents

To contextualize the DSA's approach, we trace the regulatory evolution governing online platforms and some of the struggles in addressing complex, international disinformation campaigns. The EU, in fact, had an initial reliance on self-regulation, which exhibited flaws potentially enabling threats to proliferate. Over time, binding rules with oversight

mechanisms emerged, although balancing such a complex topic as disinformation against issues of virality, speech and evidentiary burdens remained challenging.

The *E-Commerce Directive* (ECD) promulgated in 2000 formed the EU's inaugural horizontal framework embracing the Web 2.0 promise, by limiting platforms' liability for third-party illegal content under notice-and-action procedures [56]. Jurisdictional arbitrage was permitted under the country-of-origin principle allowing platforms to forum shop across Member States for lightest oversight by solely adhering to headquarters' laws [109]. The ECD's broad safe harbor exemptions also granted platforms expansive discretion in setting inconsistent, opaque content rules interpreted as permitting removal of legal speech despite due process gaps [102]. And self-regulatory reliance critically meant limited transparency, evidentiary burdens or accountability around algorithmic amplification of illegal and emotionally charged polarizing content [170]. Thereby the ECD arguably established conditions allowing contemporary disinformation ecosystems exploiting psychological biases to emerge through selective virality-maximizing curation without recourse.

The 2018 EU *Code of Practice on Disinformation* (CPD) represented a first acknowledgment of previous regulatory failures, attempting to secure voluntary commitments from prominent platforms including Facebook, X (Twitter), and Google to limit coordinated influence operations through fact-checking partnerships, content demotion and demonetization [46]. Introduced as a bid to combat fake news amidst the upcoming 2019 European Parliament elections, the Code was an important first step in aligning different stakeholders to set standards on fighting disinformation. Yet, its non-binding nature and processes beholden to commercial viability attracted extensive critiques [30, 34, 48]. Following an internal Assessment of the Code, the Commission itself issued a Guidance in May 2021 underlining how it should be improved in areas such as definitions, clearer procedures, imprecision around prohibited behaviours and lack of monitoring mechanisms [47, 48]. Self-assessment constrained transparency among data sharing restrictions to protect trade secrets and privacy [29]. Moreover, oversight depended on irregular corporate disclosures lacking standardization. Thereby while acknowledging scale of threats, its remedies centered platforms rather than the impacted public.

In response to some of the criticism faced, the Commission therefore introduced a strengthened and updated Code in 2022, expanding on previous partnerships by requiring signatories document procedures, data access protocols and effectiveness indicators [49]. Stricter identity verification aims reducing influence operations and synthetic media usage, while efforts are underway to balance counter-speech quality with the plurality of views (as reflected in the DSA through DD and NTD). However, with expanded data gathering duties arising alongside privacy sensitivities after General Data Protection Regulation (GDPR) [45] adoption, persistent transparency—accountability tradeoffs remain negotiated through non-binding codes subordinate to commercial constraints [2].

So in effect, the EU trajectory that informed the DSA represents gradual acknowledgment of self-governance limitations, giving way to formal cooperation against complex threats spanning borders. However, the challenge lies in reconciling binding regulations with the evidence required for accuracy, oversight, and proportionality, amidst pressing calls for action. The binding nature of the DSA significantly broadens obligations, aiming to strike a balance in addressing these challenges.

3.2 Limitations and persistent gaps in the DSA

The DSA adopted in 2022 represents a significant evolution from the self-regulatory Code to a comprehensive legal framework for regulating platforms' content moderation practices, with a comparison summarized below in Table 2.[3]

While the Digital Services Act represents progress beyond the self-regulatory efforts previously seen, scholars warned about some of its persistent gaps, inconsistencies and unexamined assumptions [78, 82, 104, 109, 169]. We now present an in-depth legal and technical critique of the DSA along with psychological, behavioral and ethical perspectives to enable a holistic, evidence-based assessment.[4]

3.2.1 Legal and technical critique

As the analysis of the ECD and 2018's CPD revealed, persistent gaps remain regarding inconsistent definitional precision, uneven transparency and accountability mechanisms, constrained oversight across platforms, and limited harmonization in responsibilities and redress options [102, 170]. The subsequent legal and technical critique examines precisely how such issues manifest as lingering ambiguities, discretionary authority and procedural weaknesses within the DSA framework itself.

[3] The table compares the E-Commerce Directive, EU Code of Practice on Disinformation, and Digital Services Act across: legal basis, disinformation definitions, accountability, psychological factors, algorithm transparency, freedom of expression safeguards, redress options, oversight scope, and ethical considerations. It highlights the evolution of EU regulations addressing online disinformation, preparing for an in-depth analysis of the DSA.

[4] Hereby, the categorization and analysis of critiques was developed through a comprehensive qualitative coding methodology based on grounded theory principles [97]. Our Appendix 3.2 provides further details on the coding methodology and development process for the taxonomy structure categorized into legal, technical, psychological, behavioral and ethical critiques.

Table 2 Comparative analysis of regulatory frameworks for disinformation

Criteria	E-Commerce Directive (2000)	EU Code of Practice (2018)	EU Strength. Code of Prac. (2022)	Digital Services Act (2022)
Legal Basis	Directive under EU Law	Voluntary Agreement	Voluntary Agreement	Regulation under EU Law
Definitional Precision	Lacks precise definitions on disinformation	Circular, expansive definitions	Expanded provisions on prohibited tactics, techniques and procedures	Attempts precision but retains ambiguities
Accountability Mechanisms	Self-Regulatory, Limited Oversight	Self-Assessment, Non-Standardized Reporting	Service-level indicators introduced	Mandated Audits, Reporting Obligations
Algorithmic Transparency	Not Addressed	Opaque Policies	Commitments to limit manipulative AI systems	Some Oversight but Limited
Freedom of Expression Safeguards	Not Explicit	Lacks Precision	Acknowledges the right of freedom of expression	Attempts Safeguards but Incomplete
Redress Mechanisms	Not Explicit	Inadequate	Appeal process introduced	Introduces Complaint Processes
Scope of Oversight	Limited to EU member state of origin	Broad but non-binding	Broad but non-binding	EU-wide Harmonization
Psychological Factors	Not Addressed	Not Addressed	Encourage media literacy, recommender system transparency	Partially Addressed; Lacks depth on cognitive biases
Ethical Considerations	Not Addressed	Not Addressed	Not Addressed	Ethical Risk Assessments Encouraged

Definitional Ambiguities Open questions around what constitutes a "*systemic risk*" under the DSA (Code: RA) stand [169]. Views diverge on "whether a systemic risk is one that has an impact on a system and, if so, which system(s), or if it is a risk that is caused or exacerbated by a system" [82]. This ambiguity around the core notion of systemic risk creates implementation challenges for platforms undertaking assessments, especially given the availability of public DSA readiness checklists [173]. Alongside that, the definition of "*harmful*" content tackled through notice-and-takedown procedures (Code: NTD) in Articles 14-15 lacks binding legal force and grants excessive discretion to platforms over concepts like intent, truth and harm without adequate precision (Article 26) [109]. Building on those definitional ambiguities, provisions emphasize metaphorical definition of "*virality*" instead of granular technical factors like source credibility, frequency, recency, and message exposure that psychologically enable belief formation (Articles 26, 27) [78].

Uneven Obligations Differentiated responsibilities across online intermediaries covered under due diligence provisions (Code: DD) in Articles 10-15, 35-37 do not sufficiently account for varying capacities and risks of enabling harms, disregarding potential anti-competitive impacts that contradict intended proportionality. As denounced by [107], VLOPs might be able to 'capture' and influence their mandatory auditors. As remarked, VLOPs must undergo yearly independent audits (Code: AD), yet because there are limited platforms that meet the VLOP threshold, these platforms might dominate the demand for audits. Auditors may then feel pressure to avoid overly scrutinizing VLOPs in order to retain their business, creating a 'go-to' choices of auditors for big platforms. The ratio of increasing platform user threshold and heightened obligations might overlook risks from smaller platforms [107]. Even if they have fewer users, smaller platforms can still propagate disinformation or enabling discrimination [153]. So the VLOP categorization seems limited from a risk perspective, and may allow harmful practices to go undetected on smaller platforms or other private channels.

Redress Limitations Absence of minimum harmonized standards for notice-and-action procedures (Code: NTD) in managing content removal disputes risks inconsistent, arbitrary handling of misinformation reports across Member States (Article 14) [169]. Redress mechanisms introduce complaint processes but lack impartial review prior to speech restrictions, disregarding horizontal effect principles (Articles 17, 18) [107]. As noted by [78], content moderation rules of the DSA—notice and takedown, trusted flaggers, dispute resolution, risk management systems—should be selectively expanded to apply directly to Generative AI developers and deployers. This might contribute in having problematic content to be flagged and blocked at the source, before proliferation.

Monitoring Barriers No transparency requirements or oversight mechanisms exist for demotion of legal content, granting unchecked power to dominant platforms over online discourse flows (Article 27 & Recital 56).[5] Expanded reporting and auditing enhance accountability but standardization gaps and data access limits may constrain holistic monitoring by authorities and researchers (Articles 13, 28, 35) [107, 183]. In particular for auditing, audit findings may inform future regulatory interventions or enforcement actions if systemic non-compliance is uncovered (Code: AD). Nevertheless, audits are limited to assessing procedural compliance rather than evaluating substantive fairness, proportionality or effectiveness (Recital 92), also in regards to AI systems deployed [28]. As remarked before, the level of genuine independence between powerful platforms and auditors remains unclear (Article 37).

Shadow Banning Connected to transparency regulation in content moderation, the DSA is the first to address the issue of shadow banning directly. By defining it in Recital 55 and outlining specific provisions in Articles 14 and 17, a foundation for identifying and addressing this controversial practice is provided. The regulation aims to hold platforms accountable for their content moderation actions and enhance transparency by providing users with information about why certain content is affected [157]. Yet the DSA adopts a particular definition of shadow banning referring to accounts being secretely suspended, rather than the more modern interpretation of secret visibility reductions, such as lower social media posts viewing [108]. This lack of adoption of a more up-to-date interpretation of shadow-banning ought to be kept in mind, from a legal standpoint, while differentiating different types of visibility reductions. Moreover, the complexities surrounding the detection and proof of shadow banning practices may make it more challenging to enforce than other forms of content moderation, such as tackling false information.

3.2.2 Psychological and behavioral critique

The traversal of preceding EU efforts exposed minimal incorporation of insights from psychology and behavioral science into countering mechanisms enabling real-world disinformation proliferation. The non-binding CPD lacked obligations around transparent algorithmic systems that exploit engagement and vulnerabilities, while the ECD enabled opaque amplification of emotionally charged content to continue unchecked [170]. This risks perpetuating blindspots around the outsized influence of partisan biases, opaque algorithmic harms from engagement-maximizing systems, and emotional appeals enabling the spread of disinformation [41, 78].

Bias and Identity Its reliance on post-hoc content corrections and debunking through notice-and-takedown procedures (Code: NTD) underestimates robust behavioral evidence revealing the outsized influence of preemptive friction nudges counteracting confirmation tendencies [94, 147], motivated reasoning [137], or identity-protective cognition [18] through induced self-reflection rather than external dispute (Article 27). Provisions emphasize metaphorical notions of "*virality*" instead of granular technical factors like source credibility and message exposure frequencies psychologically enabling belief formation (Articles 26-27, Codes: NTD, RA) [63, 64, 78].

Credibility and Engagement Additionally, oversight limitations enable engagement-maximizing ranking systems to exploit cognitive capacity gaps constraining users' systematic evaluation of information credibility when exposed to disproportionate volumes of borderline unreliable content [24] (Article 29, Code: RA). This propensity risks enabling various contextual and demographic inclusion/exclusion harms [61] from opacity. Reliance on platform counter-speech risks backfire effects when ideologically kindred content is tagged as misinformation, entrenching beliefs [67] (Article 27). This effect might scale if not recognizing long-standing scholarship on selective exposure and homophily, thereby limiting breeding epistemic distrust of institutional fact checkers in polarization contexts [6, 144].

Psychological Depth and Targeting More broadly, insufficient attention to established emotional, moral intuition and identity-adjacent appeals giving disinformation intrinsic force overlooks extensive models clarifying the depth of human vulnerability beyond narrow information deficits or viral harm frames [41] (Article 26, Code: RA). In terms of targeting risks, the lack of credible, scalable technical and procedural safeguards against behavioral microtargeting risks enabling manipulative disinformation campaigns (Articles 24, 29-30, Codes: DD, AT, RA). Since the DSA doesn't outright ban certain ad targeting or recommender practices, determining violations will require case-by-case analysis [78, 107]. Lastly, still related to design choices maximizing engagement, addictive design features have raised alarm also for Members of the EU Parliament (MEPs) who voiced their concerns and called for a ban on addictive techniques of recommender systems [55]. The DSA should clarify and reinforce risk assessment and mitigation requirements imposed on VLOPs concerning these aspects, as outlined in Articles 34 and 35.

[5] As a clarification, recitals shall serve to elaborate Article objectives without constituting obligations. Thereby their role is unambiguously scoped for reader understanding. Within this perspective, Recital 56 reproves particulary useful while discussing monitoring barriers on platform that could *"set the rules of the game, without effectively identifying and mitigating the risks and the societal and economic harm they can cause"*.

3.2.3 Ethical and philosophical critique

As highlighted during the analysis of regulatory antecedents, prior self-governance centered authority within commercial entities rather than democratically accountable institutions and citizens, raising legitimacy concerns [109]. For instance, the ECD's intermediary liability approach granted platforms control over speech removal while limiting avenues for public participation in content governance. The ethical critique explores more abstract yet fundamental tensions between privatized speech regulation and participatory self-determination perpetuated under the DSA, alongside duties around truth and transparency.

Private Governance and Participation Delegating speech regulation to private platforms (Code: NTD) conflicts with commitments to participatory self-governance, instead expanding corporate control over discourse aligned with private interests rather than civic oversight [8, 17, 29, 135]. This questions the legitimacy of privatized judgments over truth and risks disregarding expressive diversity. Absence of checks on algorithmic demotion (Code: RA) elevates private economic calculations over considerations of expressive diversity and exposure plurality. This prioritizes efficiency over ethics. Oversight and due process are needed for any restrictions on legal speech [96, 161].

Truth Judgments and Shared Liability Compelled counterspeech and warnings on otherwise legal content creates tensions with liberties of conscience and freedom of uncoerced thought. Safeguards should be strengthened for only employing this extra speech as a last resort based on judicial review [31, 181]. Similarly, restrictions based on opaque judgements of truthfulness and potential virality, rather than direct incitement standards, contradict assumptions behind epistemic rights and the "marketplace of ideas." In this sense, generative AI solutions—such as OpenAI's ChatGPT [158, 159, 166]—yet would not fit into the "hosting service providers" since they generate content themselves rather than just storing and disseminating user-generate content. Therefore, loopholes in protections persist regarding risks from harmful AI-generated content posted publicly that cannot be fully removed once spread (Code: NTD) [78]. Also, provisions take an individualized view of viral harms rather than shared duties of good-faith public discourse. The DSA rules only apply (Code: DD) to "online platforms" that disseminate information to the general public. Private messaging services like WhatsApp or Telegram groups are excluded. But harmful content generated at scale, or by polarized groups, could easily spread in these private channels outside the DSA's scope [78]. Thus, a solely liability-driven approach might underestimate collective social responsibilities around ethical speech cultivation.

Allocating Moderation Resources Under a geopolitical lens, a possible unintended impact for moderation activities is the "Brussels effect" whereby resources are disproportionately allocated to mitigating EU-specific risks—likely identified in assessments (Code: RA) or gauged by EU public opinion—negatively affecting content moderation practices elsewhere globally. This underscores needs for research into differential impacts of the DSA's provisions across regions [78, 169].

Lack of Direct Journalist Protections The DSA introduces key provisions aimed at improving online safety also for the Media and Journalism sector. These provisions (Code: NTD, DD) empower trusted flaggers to identify and address illegal content (Article 22), allow national authorities to issue orders for the removal of harmful content (Article 9), and establish effective mechanisms for content notification and action (Article 16). The risk governance system as laid down in Articles 34, 35 and 37 could play a part in the protection of journalistic and news media content. While the DSA provides tools for setting the ground for protecting freedom of expression and pluralism of media by mandating that VLOPs identify systemic risks, it does not specifically target the protection of journalists and media professionals [22, 54, 136], despite acknowledging the importance media freedom as reported in Recital 47 and 81.

4 Research approaches to evaluate DSA impact on disinformation

Having analyzed multi-dimensional critiques revealing meaningful gaps in the DSA's approach, undertaking comprehensive empirical research should assess its real-world efficacy across involved stakeholders. Specifically, the legal analysis flagged ambiguities in key definitions and inconsistencies in obligations across platforms. The technical review highlighted monitoring barriers, while the psychological critique called out assumptions on how false content spreads. The ethical lens questioned privatized speech regulation. To assess such efficacy, research should trace impacts back to the DSA's core mechanisms. For example, will notice-and-takedown processes consistently and accurately curb illegal content? How will restrictions on targeting ads affect disinformation campaigns? Do risk assessments lead to meaningful change?

This section gives an overview of quantitative and qualitative methods to start answering such questions. Strategically combining methods will allow assessing if the DSA achieves its aims around transparency, accountability and empowering users.

4.1 Overview of research needs and objectives

The current goal is to produce detailed evidence on the DSA's performance and areas for improvement, enabling

systematic enhancements to better achieve its aims of protecting the public. To do so, it's vital to ensure assessment methods directly related to DSA's operational mechanisms, aiding in strategic updates that are informed by solid, actionable insights throughout its deployment.

We define three main types of methods. Quantitative techniques facilitate assessing phase-in outcomes related to advertising transparency, policy impacts on moderation consistency, and tracing notice-and-takedown efficacy (Code: AT, NTD, DD). Qualitative analyses provide equally vital contextualization regarding governance adaptations during high-risk events and platform discourses on evolving responsibilities (Code: AD, RA). Multi-disciplinary methods spanning computational analytics, focused audits, ethnographic observations and surveys can construct nuanced evaluations across involved stakeholders.

4.1.1 Quantitative research

Content Analysis Studies Longitudinal content analysis enables evaluating utilization and efficacy of notice-and-takedown procedures (Code: NTD) under Article 14 by tracking volumes and categories of disinformation removed over time. Segmenting takedowns by dimensions like platform, content type, source, and response rates provides granular impact metrics aligned with transparency objectives around illegal content reporting workflows [130]. Advanced text classification techniques can categorize removed content by type, topic, sentiment to provide granular volumetric impact metrics [44]. To audit improvements in advertising transparency under Articles 24 and 30 (Code: AT), content analytics can categorize disclosed ads in mandated repositories by targets, sponsors, geographic origin, topics, and media types. Related studies have analyzed voluntary ad archives to uncover transparency gaps [33, 140]. While to examine platform policy evolutions after Article 27 (Code: RA), text analysis can track changes in community standards and terms of service on mis/disinformation before vs after DSA implementation. Semantic modeling can detect subtle policy shifts across platforms over time, as witnessed by similar work analyzing platforms' policy dynamics around emerging speech issues [118, 143].

Independent Audit Studies Audit studies[6] enable causal assessment of DSA impacts by directly testing platform responses through carefully controlled researcher-instigated interventions [87, 172]. Computational tools allow efficient large-scale auditing of critical areas like ad transparency, fake accounts and disinformation detection [88].

To assess impacts of diligence obligations (Code: DD), researcher-instigated audit studies can systematically post synthetic disinformation from control vs treatment groups, and compare detection and removal rates based on platform size and due diligence status. Factorial experimental designs can isolate impacts of specific obligations. Similar audits have assessed content moderation consistency across platforms [148]. In a similar vein, to measure AT improvements under Articles 24 and 30, audits can evaluate completeness, formats, and accuracy of political ad disclosures before vs after DSA roll-out [163].

Computational Analysis Advanced computational techniques like machine learning and network analysis allow scalable unstructured data investigations to evaluate impacts of critical DSA provisions. For instance, natural language classifiers enabled by neural models can audit samples of platform content containing warning flags or misinformation markings before and after implementation to quantify accuracy, truthfulness and contextual alignment[106]. This tackles identified criticisms around transparency and proportionality of counter-measures under Article 27 [38, 43, 68, 180]. Separately, temporal analyses tracking toxicity trends or categorizing discourse by topics using sentiment modeling and dynamic topic classifiers provide vital indicators of ecosystem impacts resulting from risk mitigation efforts under Article 26 [119, 145]. This helps examine efficacy in limiting systemic threats from viral disinformation based on content flows at population scale [79, 89, 174]. Lastly, network analysis enables mapping increased connections and centralization patterns around public-private collaborations tackling disinformation issues following DSA adoption, revealing integration of oversight duties under Article 31 [76, 98, 134].

Statistical Modeling Statistical techniques like regressions, interrupted time series analysis and multidimensional panel data modeling allow robust quantitative examination of DSA's impacts on complex phenomena including misinformation diffusion dynamics and fact-checking effectiveness across diverse platforms, issues and time periods [187]. These approaches enable detecting signals within large-scale observational datasets that complement field experiments.

For example, segmented regression combined with quasi-experimental difference-in-differences analysis facilitates comparing misinformation spread trends before and after DSA adoption between control and treatment groups of platforms based on characteristics like resource levels and growth stages. This reveals policy efficiency by associating implementation timelines with campaign diffusion shifts under Article 26 (Code: RA). Applying machine learning classifiers on content provides prediction improvements [93, 99, 115]. Separately, interrupted time series tracking engagement metrics with fact-checks/verified data spanning DSA roll-out can associate significant changes in dimensions

[6] NB. Hereby, research *audit* studies as used here refer to independent experiments by researchers, distinct from mandated DSA compliance audits in Article 37 and the Delegated Act [50].

like exposure and recall with efforts promoting authoritative information under Article 27 (Code: NTD). Vector autoregression strengthens attribution by accounting for metric interdependencies. Panel structures with platform, issue and event segmentation enable generalization checks. Backdrop causal inference algorithms further bound claims for observational studies [15, 91, 101, 114, 176].

4.1.2 Qualitative research

Interviews and Focus Groups Interviews and focus groups generate vital subjective insights into on-the-ground realities shaped by formal DSA mechanisms and lived implementation experiences [100], as well as assessing governance reliance on internal or public readiness checklists [173] and evaluate perceived comprehension of obligations. Semi-structured interviews with internal governance teams could reveal challenges in operationalizing systemic risk assessments (Code: RA) across different services [169], allowing evaluation of provisions like standardized reporting flows (Code: DD) in Article 31 and participation in risk analysis (Code: RA) under Article 26. Focus groups surveying external citizen attitudes pre/post adoption facilitate gauging perceived impacts on information quality and algorithmic transparency (Code: AT) tied to Articles 24 and 29-30. This provides multi-stakeholder contextualization of shifts attributable to key provisions. Specific questions could also be directed towards trusted flaggers under Article 15, examining experiences submitting notices (Code: NTD) and interacting with evolving notice-and-takedown platforms. Polling similarly enables quantifying user perceptions on experiences with online accountability measures [69].

Ethnographic Observation Interviews and embedded analysis within digital oversight bodies and technology firms aids directly examining operational adaptations arising from DSA due diligence obligations (Codes: NTD, DD, RA). When well-resourced, these initiatives provide vital empirical findings revealing frontline policy implementation obstacles, while also supporting data assessments and enforcement actions under Article 40 [90, 111].

Equally important, sustained participant observation across product operations teams facilitates identifying crucial changes in governance processes made towards realigning company norms on transparency (Code: AT), risk analysis rigor and accountability towards regulatory aims amidst uncertainty. For instance, directly observing high-risk policy units over time would uncover ground realities, best practices and persistent friction points that shape how formal rules translate into practice (Code: NTD, RA), complementing external computational impact evaluations [125].

Closely involved research gives essential methods for studying technology platforms [3, 124].

Discourse and Narrative Analysis Discourse analysis enables tracing subtle shifts in public narratives revealing underlying logics guiding platforms' cultural positions on emerging responsibilities, by combining computational techniques with critical reading (Code: AT) [71, 83]. For instance, analysis over time could determine evolving rhetoric usage in community standards (Code: NTD) and risk governance priorities (Code: RA) before/after DSA mechanisms enter force under transparency articles (Code: AT) and recurring assessments (Code: RA). For example, semantic analysis [71, 83] and opinion detection [4] of community standards along with transparency reports [175] can surface evolving constructions of concepts like misinformation, platform responsibilities, and approaches to risk mitigation. Comparative approaches across firms also elucidate differences in translating common pressures, be that from broader public discourse expectations or specific regulatory provisions [3].

Comparative Case Studies In-depth case examinations enable contrasting context-specific translations of formal DSA duties (Codes: NTD, AT, DD) into ground-level content moderation and risk governance practices by platforms across elections, health issues and geopolitical events (Code: RA) before/after adoption [5, 44, 149]. For instance, comparative analysis of policy adjustments (Code: NTD), collaborations with authorities (Code: DD), communications norms (Code: AT) and computational detection systems (Code: RA) around events can surface crucial commonalities and differences attributable to phase-in impacts [86, 92]. Detailed analyses of content moderation approaches [93], misinformation detection [184], collaborations with authorities [86], and communications around individual events [70] can reveal commonalities and differences shaped by regulatory pressures. This way, case studies can typologize and categorize interventions across events to identify process changes attributable to DSA mechanisms like notice-and-takedown procedures (Article 14), advertising transparency provisions (Article 24), or coordinated oversight under the European Board for Digital Services (Article 47), as performed by similar comparative case studies [86, 152].

4.1.3 Mixed methods research

Strategically combining quantitative and qualitative methods through careful integration and triangulation allows developing robust, multi-faceted understandings of the DSA's complex impacts. As [112] articulate, mixed methods provide strengths including corroborating findings across different data types, capturing nuances, and integrating broad trends from computational analytics with contextual qualitative insights. A preliminary overview of methodological approaches and sources is offered in Table 3.

Table 3 Methodological matrix for investigating DSA's impact on platform governance

Dimension of analysis	Quantitative approaches	Qualitative approaches	Potential data sources
Perception and Experience	Surveys	Semi-struct. Interviews, Focus Groups	Policy makers, Governance Teams, Public
Publicly Stated Policies	Content Analysis of Official Documents	Discourse and Narrative Analysis	Transparency Reports, Community Guidelines
Organizational Processes and Norms	Descriptive Statistics, Structural Equation Modeling	Ethnographic Observation	Internal Reports, Direct Observations
Platform Interventions in High-Risk Events	Event Study, Logistic Regression	Comparative Case Studies	Platform Logs, Press Releases
Compliance and Accountability Mechanisms	Compliance Audits, Descriptive Statistics	Expert Reviews, Stakeholder Interviews	Compliance Reports, Legal Interviews
Perception of Systemic Risks	Factor Analysis, Multivariate Regression	Semi-Structured Interviews with Internal Stakeholders	Risk Assessment Reports, Internal Interviews
Multi-Stakeholder Collaboration	Network Analysis, Social Network Metrics	Stakeholder Interviews, Focus Groups	Collaboration Agreements, Meeting Minutes
Public Perception of Information Quality	Sentiment Analysis, Time-Series Analysis	Focus Groups with Citizens	User Reviews, Public Surveys
Internal Decision-Making	Decision Tree Analysis, Random Forest	Ethnographic Observations within Organizations	Decision Logs, Direct Observations
Day-to-Day Operational Changes	Change-Point Analysis	Longitudinal Ethnographic Observation	Change Logs, Direct Observations

As already acknowledged, surveys distributed to platform governance teams, civil society groups, journalists and citizens can gauge expert and public perceptions of online information quality changes (e.g. due to transparency improvements under Article 13 and 24). Combining the survey results with field and lab experiments that directly manipulate exposure to misinformation before vs. after DSA roll-out can help quantify impacts on recipients while also capturing subjective attitudes and experiences [139, 185]. Surveying the European Board for Digital Services per Recital 134 could provide additional insights. For instance, experiments exposing groups to misinformation across platforms before and after DSA adoption, coupled with surveyed perceptions of its prevalence, would enable richer assessment.

Computational analyses like dynamic computational models quantifying the diffusion patterns of known disinformation narratives across platforms and time periods can be enriched through multi-faceted qualitative data [122, 138]. Specifically, discourse analysis of shifting platform policies and narratives around misinformation can surface logics shaping content interventions [110]. Ethnographic observations within platform governance teams examining changes in day-to-day content moderation practices and norms explicate on-the-ground contexts underlying computational trends [105].

Triangulating large-scale quantitative indicators tracking activities like volumes of content removed via notice-and-takedown procedures with qualitative interviews tracing platforms' procedural shifts and implementation challenges might allows cross-validating findings across methods [74]. Similarly, in-depth comparative case study analyses of platform responses across risky societal events before vs. after DSA adoption can be reinforced by interrupted time series statistical models estimating misinformation diffusion trends [19].

Careful coordination in framing inquiries, timing data collection, and integrating computational and qualitative approaches enables producing robust, multidimensional insights. As reported by [178], following established guidelines for mixed methods research design is key to effectively triangulate findings across approaches when evaluating complex policy interventions like the DSA.

To summarize the current section, Fig. 1 maps between research approaches, goals, and stakeholders for a comprehensive understanding of the research landscape in DSA studies. The figure was obtained through a methodological framework exposed in Appendix 3.1.

4.2 Methodological challenges and limitations

The proposed research agenda aims to enable multifaceted assessment of the DSA's impacts across stakeholders. Yet realizing this potential depends on overcoming methodological challenges that shall be acknowledge. Specifically, the extensive legal, technical, psychological and ethical critiques surface open questions regarding the DSA's real-world effects that the outlined quantitative and qualitative methods intend to address. To achieve that, we first need to recognize possible difficulties isolating the DSA's influence amidst adjacent regulations, constrained

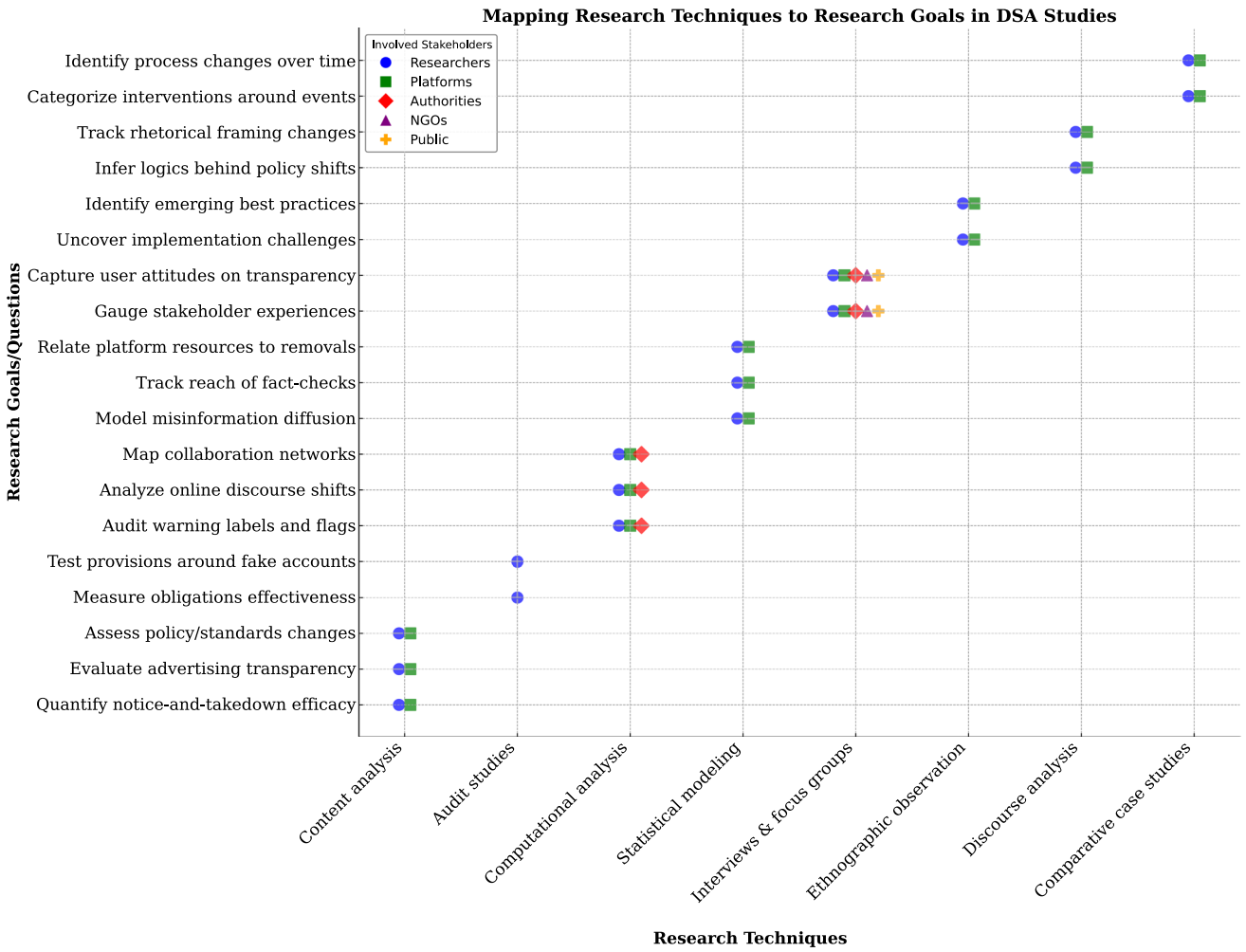


Fig. 1 Mapping of Research Approaches to Research Goals and Stakeholders in DSA Studies. The graph is color-coded to indicate involved stakeholders, and points are offset for readability when multiple stakeholders are involved for a single goal

cross-context generalization, and dependence on voluntary platform data sharing.

Touch-points and Conflicts in the EU Legislative Ecosystem In Sect. 2.2 we briefly touched upon the AI Act's provisions explicitly referring to the DSA (i.e., recital 60u, 60aa, 60v, 70d). Nevertheless, to keep the current research scope self-contained we had to exclude from the current analysis secondary EU regulations pertaining to digital data and services. Yet, it should be acknowledged that the interplay between the DSA and other concurrent or forthcoming EU regulations directed at digital platforms presents challenges in isolating the specific effects attributable to the DSA's provisions. For instance, the *AI Act* [37] impose obligations around algorithmic transparency and provisions (especially Articles 13 and 14) against manipulative AI systems that intersect with the DSA's focus on mitigating algorithmically amplified disinformation [62, 78]. To delineate these relations, we report here the main regulations likely to impact the DSA in terms of platforms and data governance.

- The *General Data Protection Regulation* (GDPR) [45], in force since 2018, establishes horizontal rules governing the processing and free movement of personal data in the EU. Its key principles include lawfulness, fairness and transparency; purpose limitation; data minimization; accuracy; storage limitation; and accountability. The regulation applies broadly to organizations processing EU residents' personal data. In terms of connections, the GDPR and DSA potentially align on transparency duties around data processing activities. Expanded platform content moderation duties under the DSA Article 14 may conflict with GDPR data minimization requirements (Article 5), if applied excessively. There is also a risk of tension regarding compliance burdens, as meeting DSA's additional transparency and reporting obligations could

create challenges for simultaneously meeting GDPR's data protection impact assessment and breach notification requirements. Ongoing coordination between national DPAs and the European Board for Digital Services (DSA Article 47) will help reconcile such potential tensions, ensuring cohesive interpretation and mutually reinforcing accountability aims while safeguarding innovation.

- The *Digital Markets Act* (DMA) [59] entered into force in 2022 alongside the DSA. It aims to ensure fair competition and contestability for core platform services such as search engines, social networks, and marketplaces. Specifically, the DMA targets large gatekeeper platforms with entrenched market power, imposing rules against self-preferencing, leveraging data advantages, and limiting interoperability and multi-homing.[7] Potential clashes could arise regarding consistency in gatekeeper designations if thresholds or criteria diverge over time. Similarly, varying transparency duties and restrictions around data usage and self-preferencing might create conflicting platform obligations under each regulation. However, the regulations also align on numerous dimensions, for instance reinforcing interoperability duties and limiting behaviors constraining user choice.
- The *Data Governance Act* (DGA) [60] passed in 2022 and applicable in September 2023, aims to facilitate data sharing in the EU by enabling data intermediaries and data altruism frameworks. It creates a notification system for data intermediation service providers and allows entities that collect data for general interest purposes to register as "data altruism organizations." In terms of scope, the DGA focuses on enabling voluntary data flows. With the DSA, potential alignments include facilitating external researcher access to platform data: indeed, DSA Article 31 mandates data access for studying risks, while DGA Chapter 2 governs data sharing and re-use from public sector bodies. However, tensions could emerge around intermediary liability if DGA data sharing provisions expand platforms' accountability for user-generated content. And DGA recognition of data cooperatives and altruism efforts could create unclear overlaps with DSA compliance.
- The *Data Act* [58], proposed in 2022 and entered into force on 11 January 2023, establishes harmonized rules on data access and use (1.) between businesses; (2.) between businesses and consumers; (3.) and between businesses and the public sector. The new measures complement the DGA, yet the former regulates processes and structures that facilitate voluntary data sharing, while the Data Act clarifies rights and usage conditions regarding data, sustaining incentives for data holders to continue investing in high-quality data generation. In terms of touchpoints with the DSA, the Data Act's Chapter II on business-to-consumer and business-to-business data sharing complements the DSA's focus on transparency and access for researchers. Yet tensions could emerge around expanded data access if the Data Act's provisions extend platforms' content moderation duties or notice-and-takedown responsibilities under the DSA Article 14.

Consistency among Member States Despite the theoretical discussion upheld throughout this text, we yet shall be cognizant that difficulties might arise in generalizing research findings across the sociotechnical diversity of platforms and national contexts within the EU's Member States. Individual platforms have distinct technical affordances, algorithmic architectures, business models and user demographics that shape how disinformation manifests and spreads [6, 36, 77]. Additionally, the establishment of national Digital Service Coordinators (DSCs) under Article 38 of the DSA, tasked with national-level supervision and enforcement, may lead to divergences in oversight approaches based on resources, priorities, and local legal contexts across Member States. For instance, DSCs in Western Europe may devote more resources towards mitigating foreign disinformation around elections, while DSCs in Eastern Europe might prioritize e.g., monitoring domestic disinformation around social issues [65, 66, 85]. Analyzing interactions between national DSCs and VLOPs can also surface crucial nuances in oversight approaches across Member States' sociopolitical contexts. This specificity necessitates caution in making broad generalizations; yet, analyses can inform coordinated efforts towards harmonization of responsibilities under the DSA across diverse environments through:

- Systematized sharing of oversight techniques and campaign responses between national Digital Service Coordinators to highlight best practices through the European Board for Digital Services.
- Paired auditing and interview studies simultaneously assessing platform and DSC adaptation across Member States to highlight inconsistencies.

Instead, comparative research highlighting differential manifestations and impacts of disinformation, as well as variations in platforms' responses, will be essential to account for key distinctions across services and countries [141, 162].

Obtaining Research Data Obtaining adequate access to platform data represents a major constraint, as transparency

[7] While the DSA and DMA share complementary aims around ensuring online platform accountability, their scopes differ. The DSA establishes horizontal rules governing illegal content, platform responsibilities, and algorithmic transparency. In contrast, the DMA intervenes in platform competition dynamics and economic fairness issues through a set of do's and don'ts that only apply to designated gatekeeper entities meeting specific thresholds.

provisions in the DSA grant only limited data sharing with researchers (Articles 31 and 40). Researchers may need approval through vetting procedures per Recital 97, further constraining access. This risks selection biases if computational methods rely solely on voluntary self-disclosure of cherry-picked data by platforms. Over time, major platforms have increasingly restricted third-party researcher access to data through API limitations like Facebook's Ad Library and CrowdTangle analytics tools in 2018, exerting control over data access, limiting transparency, and hampering independent research in the public interest [21, 120].

The DSA introduces welcome but restrained provisions allowing vetted researchers to access data from VLOPs to analyze systemic risks under Article 40. However, the exact data types, quantities and protocols remain vague and incomplete. There are concerns that excessive corporatized control over data access could constrain the efficacy of independent DSA impact analyses [82, 109, 169]. Therefore, policy evolutions may enable appropriate access to sufficiently detailed, representative data based on the public interest merits of independent research. Data access frameworks with democratic oversight and mandatory sharing for qualifying purposes could be impactful, but little has been achieved through the updated Code of Practice on Disinformation [2, 49]. Technical solutions granting access could involve requiring availability through open APIs of privacy-preserving synthetic platform datasets, development of trusted intermediary data institutions to broker access, and heightening researcher intimacy levels to unlock voluntary sharing.

5 Conclusion

This paper aimed to catalyze scholarly discourse on empirically evaluating the DSA's efficacy in achieving its central goal of combating online harms. It provided necessary context on the DSA's mechanisms and expected effects to situate the analysis. An extensive multi-disciplinary critique then surfaced meaningful gaps requiring ongoing regulatory and evidentiary oversight. Most critically, it proposed an encompassing quantitative, qualitative and mixed methods research agenda tailored to assessing the DSA's influence on complex factors enabling disinformation. Such analysis encompassed assessing advertising transparency, content moderation consistency, risk assessment outcomes and procedural shifts in platform governance.

In scoping this agenda, the analysis transparently grappled with methodological difficulties around isolating the DSA's impacts amidst adjacent regulations, ensuring cross-context generalization and obtaining adequate data access. While not definitive, the paper sought to galvanize participatory scholarship assessing and optimizing the DSA's provisions towards effectively combating online threats. Beyond conceptual directions, pragmatic insights can inform incremental evidence-based improvements through iterative, inclusive review. Independent research quantifying impacts remains essential as oversight shifts from self-regulation to binding accountability. This proposal offered necessary scaffolds for coordinated inquiry substantiating the DSA's role in this transition while shaping future adaptations.

Appendix 1: Defining disinformation in the DSA

As defined in our footnote at page 2 in Sect. 1, throughout this research paper we referred to disinformation according to the European Democracy Action Plan's definition (EDAP) i.e., "*false or misleading content that is spread with an intention to deceive or secure economic or political gain and which may cause public harm.*" [48]. While disinformation is often used interchangeably with misinformation in the literature, the latter presents a slightly different nuance. According to the EDAP, "*misinformation is false or misleading content shared without harmful intent, though the effects can still be harmful, e.g., when people share false information with friends and family in good faith.*" The nuance between the two terms lies in the intention for dissemination, which may be challenging to determine a priori. Establishing a legal definition of disinformation also posed a challenge in the context of the Digital Services Act (DSA), as policymakers aimed to ensure the term was comprehensive enough to distinguish it without over-regulating the information ecosystem [168].

Examining prominent attempts to define "disinformation" in the European regulatory framework reveals a shift away from the criterion of intention to cause harm over time, which can be associated to the difficulty in proving intent to cause harm, as emphasized by a report of the Council of Europe [10]. We report here a Table 4 summarizing the different definitions of disinformation.

The DSA itself acknowledges the complexity surrounding defining "harmful" content and avoids providing an official definition of disinformation. However, even though the regulation does not define the term "disinformation" and neither does it explicitly mention "misinformation" in the text, it identifies combatting the spread of disinformation as one of its core objectives, as we see in Recitals 9, 84, 88, 95, 104 [57].

As also previously noted in the first footnote in Sect. 1, we highlight that the most important aspect of the various definitions of disinformation is its characterization as an umbrella term. It encompasses a variety of expressions fueled by different motivations, spread through different means, and with varying levels of impact. While the term

Table 4 Definitions of disinformation in EU reports

Year	Source	Definition
2017	*Information Disorder Report* [182]	Information that is false and deliberately created to harm a person, social group, organization or country
2018	*High-Level Expert Group (HLEG) final report* [35]	All forms of false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit
2018	*Code of Practice on Disinformation, Action Plan against Disinformation* [48]	Verifiably false or misleading information that is created, presented, and disseminated for economic gain or to intentionally deceive the public, and may cause public harm
2020	*European Democracy Action Plan* [53]	Disinformation is false or misleading content that is spread with an intention to deceive or secure economic or political gain and which may cause public harm
2021	*Guidance on Strengthening the Code of Practice* [47]	The different phenomena to be addressed, while clearly acknowledging the important differences between them. Disinformation in this sense includes disinformation in the narrow sense, misinformation, as well as information influence operations and foreign interference in the information space, including from foreign actors, where information manipulation is used with the effect of causing significant public harm
2022	*Strengthened Code of Practice on Disinformation* [49]	Adopts the definition of the European Democracy Action Plan

"disinformation" is predominantly used throughout the paper, we do not exclude the use of other terms such as "misinformation," as seen in the literature and the DSA text itself. For instance, in Recital 84the DSA mandates that VLOPs pay "*particular attention to how their services are utilized to disseminate or amplify misleading or deceptive content, including disinformation*". This broader classification of harmful content extends beyond disinformation to a more general category of misleading or deceptive content, which could therefore be seen to also include misinformation.

Appendix 2: The EU Digital Services Act framework and aims

This section provides an overview of the DSA, including its context, mechanisms, and expected impacts, to situate the need for rigorous research evaluating its efficacy.

Overview of Digital Services Act

The DSA establishes an extensive set of obligations and accountability mechanisms for online intermediaries aimed at addressing illegal and harmful content at scale. It represents a major evolution in the EU's platform governance approach beyond previous self-regulatory initiatives like the 2018 Code of Practice on Disinformation [46].

We can see how the EU Commission's 2020 Assessment of the Code of Practice [49] reveals some of its shortcomings, such as inconsistent application of the Code across platforms and Member States, together with limitations due to the self-regulatory nature of the Code and its lack of extensive coverage. Amongst these identified gaps, the DSA emerged from increasing concerns around disinformation, illegal hate speech, terrorist material, and other forms of legal but harmful content proliferating online.

Structurally, the DSA seeks to harmonize the digital single market[8] by standardizing intermediary services' responsibilities and liabilities across the EU. It provides graduated obligations proportionate to entities' size, risk-posing capacities, and role in dissemination based on categorizing intermediaries into several classes with distinct duties. The graduated obligations follow a risk-based approach, with more stringent duties on very large platforms demonstrating high risks. Through transparency, accountability and due diligence obligations, the DSA aims to ensure online platforms mitigate systemic risks emerging from their activities, design and business models.

Categorization of online entities

A core feature of the DSA is the asymmetrical obligations imposed, differentiated according to the categories of online intermediaries. We can find said intermediary services grouped into the following categories:

- **Mere conduit intermediaries**: These entities merely provide transmission of information over the internet, such as broadband providers, web hosting services, and WiFi hotspot operators. They are exempted from broad

[8] We define as *Digital Single Market* the European Union's strategy to ensure the free movement of persons, services and capital in the EU's single market in the digital age. It aims to open up digital opportunities and provide fair conditions for digital services to flourish in the EU common market, while protecting consumer and personal rights. The key piece of legislation upholding this strategy is indeed the DSA paired to other legislative documents e.g., [58–60].

Table 5 Categories of online intermediaries under the DSA and their obligations

	Mere conduit	Caching	Hosting services	Online platforms
Notice-and-takedown		X	X	X
General transparency reporting	Minimal		X	X
Ad transparency reporting				X
Internal complaint/redress mechanisms				X
External auditing				X (VLOPs/SEs)
Risk assessments				X (VLOPs/SEs)
Data access provisions			X	
Consultation with civil society				X (VLOPs/SEs)

liability under Article 3 and have minimal transparency duties under Article 13. This reflects their technical role in simply transporting data packets across networks without knowledge of content.

- **Caching intermediaries**: Entities that provide transient storage and transmission of information for the sole purpose of enhancing efficiency in onward transmission. Common examples are content delivery networks (CDNs) that cache content to reduce latency. Caching intermediaries benefit from liability exemptions under Article 4 but must expeditiously remove or disable access to cached illegal content upon obtaining knowledge of its illegal nature based on Articles 5 and 14.
- **Hosting services**: A broad category encompassing online repositories that store and disseminate various types of third-party provided information at the request of users. Examples range from cloud infrastructure services, web hosting providers, app stores, online marketplaces, collaborative economy platforms, and social media networks. Hosting services are exempted from liability for user-generated illegal content if they have no knowledge of its illegal nature and act expeditiously to remove it once made aware. They are subject to DSA transparency requirements, notice-and-takedown obligations under Articles 14 and 15, and duties to provide information to authorities under Article 31.
- **Online platforms**: A subset of hosting services that actively disseminate third-party information to the public. Examples include major social media networks, video/image sharing platforms, collaborative economy platforms with public-facing profiles, app stores with browseable interfaces, and online marketplaces. In addition to hosting service obligations, online platforms must put in place complaint and redress mechanisms under Articles 17 and 18 as well as publish transparency reports under Article 13. Certain online platforms may also undergo risk assessments under Article 26 and have additional opt-in transparency duties regarding recommender systems under Article 29.
- **Very large online platforms or search engines (VLOPs or VLOSEs)**: Online platforms or search engines reaching an EU-wide threshold of over 10% of the EU population - set at around 45 million active monthly users. VLOPs such as Meta, YouTube, X, and TikTok qualify under Article 25 [51, 52]. They are subject to the most extensive DSA obligations, including external risk auditing under Article 28, data access provisions under Article 31, heightened due diligence duties under Article 27, additional transparency requirements under Article 24, and potential blacklisting under Article 27 in case of systematic non-compliance.

This graduated approach reflects the consensus that regulatory burdens should be calibrated to intermediaries' capacity and risk-influencing role, following the principle of proportionality. Indeed, one of the DSA's core objectives is to balance the power structures in the digital economy, limiting the democracy-threatening power of private companies (while not entering the realm of competition law). It aims to strike a balance between imposing adequate responsibilities on influential entities while not stifling innovation through excessive burdens on small players. To address these inherent tensions, the regulation adopts an asymmetrical approach, which has previously been adopted in EU legislation, such as in the General Data Protection Regulation (GDPR) [45]. To effectively enforce these asymmetrical obligations, it is crucial to categorize intermediaries within the DSA framework. Table 5 summarizes the key obligations for each category of intermediary service under the DSA's graduated regulatory approach.

Key mechanisms and provisions

The DSA establishes several mechanisms and provisions constituting core pillars of its regulatory approach.

As Table 6 summarizes, the core mechanisms instituted under the DSA tie to specific articles and recitals which elaborate their objectives and requirements. For instance,

Table 6 Mapping of key DSA mechanisms to related articles and recitals

Mechanism	Articles	Recitals
Notice-and-Takedown Procedures	• **Article 14** (terms need to explain content moderation) • **Article 16** (notice and action mechanisms) • **Article 17** (statement of reasons for takedowns required)	• **Recital 50** (hosting providers need notice systems) • **Recital 51** (notices should enable informed takedown decisions) • **Recital 52** (framework for notice processing) • **Recital 54** (statement of reasons requirements)
Advertising Transparency Requirements	• **Article 24** (ad transparency in platform reporting) • **Article 26** (ad transparency and disclosure to users) • **Article 39** (ad repositories for very large platforms)	• **Recital 68** (notes platforms' role in online ads) • **Recital 69** (restricts certain targeted/exploitative ads)
Risk Assessment Obligations	• **Article 26** (rules on illegal/harmful ads as a risk) • **Article 34** (systemic risk assessments) • **Article 35** (risk mitigation by very large platforms)	• **Recital 79** (very large platforms can strongly influence risks) • **Recital 80** (types of risks to assess) • **Recital 81** (consider illegal and harmful but legal content)
Due Diligence Obligations	• **Article 14** (terms transparency) • **Article 16** (notice systems) • **Article 20** (complaint systems) • **Articles 24-27** (reporting, dark patterns ban, etc.) • **Articles 34-35** (risk assessments and mitigation)	• **Recital 40** (rationale for tailored diligence obligations) • **Recital 41** (determining applicable obligations) • **Recital 47** (terms changes need to respect rights)
Independent Auditing Mechanisms	• **Article 37** (very large platforms to do annual audit)	• **Recital 92** (audits can ensure accountability) • **Recital 93** (details to be provided by Commission through delegated acts)

Article 14 and Recital 52 underpin notice-and-takedown procedures for removing illegal content, while Articles 24, 26, 30 and 39 alongside Recitals 68 and 69 impose transparency duties around online advertising.

Notice-and-takedown procedures (Art. 14,16,17) The DSA provides a harmonized notice-and-action framework for reporting and removing illegal online content across the EU. All hosting services must establish user-friendly and accessible mechanisms to allow any individual or entity to submit a notice identifying and explaining the illegality of content (Recital 52). Notices can also be sent by certified "trusted flaggers" with expertise in identifying illegal material (Article 15, Recital 61, 62), EU member state authorities, and European Commission bodies. Hosting services are also required to inform law enforcement authorities when they become aware of suspected serious criminal offenses (Article 18).

Upon receiving a notice, hosting services must expeditiously assess content's legality, remove or disable access to verified illegal material, and inform the notifier. Exact timeframes are specified based on notice source. For notices by individuals, hosting services must provide feedback about action taken without undue delay, and normally within 7 days. Removal orders from authorities must be complied with immediately, subject to judicial redress possibilities. Hosting services must also share information with national authorities regarding actions taken against illegal content (Article 31).

Additional safeguards are mandated against the submission of incorrect or unsubstantiated notices, to avoid the removal of legal content (Recital 49). Penalties for misuse can include court injunctions and financial compensation for reputational damage. Hosting services must also facilitate counter-notices by affected users disputing removal decisions, to enable users to dispute wrongful takedowns (Recital 50), and reinstate content where counter-notices are valid. Extensive transparency reporting requirements cover notice volumes, action taken, counter-notice amounts, and restored content percentages.

Advertising transparency requirements (Articles 24, 26, 39) To enable meaningful public insight into the opaque processes around targeted online advertising, Articles 24 and 30 of the DSA impose detailed transparency duties on online platforms regarding ads shown to users, specifically those related to political messaging or important public debates.

Article 30 mandates collection of trader identity and location data by platforms facilitating consumer-to-business transactions, improving traceability and accountability. In addition, Article 26 requires clear markings identifying paid ads directly in ad interfaces, as well as disclosures on

targeting parameters, enabling users to understand why they see specific ads (Recital 69). For qualifying ads, Articles 24 and 39 impose transparency duties on platforms and very large search engines to maintain publicly accessible repositories disclosing detailed information on advertisers, payments, targeting, reach, etc. in a standardized format. This aims to enable meaningful public and regulatory oversight and research into emerging systemic risks, given online platforms' and search engines' unprecedented scale and societal influence (Recitals 68, 69). Recital 68 explains the transparency rules intend to address potential harms from behavioral microtargeting like discrimination, manipulation, and disinformation amplification. Recital 26 notes requirements should not conflict with data protection laws; the focus is transparency to users, not collecting or processing more personal data.

VLOPs face additional duties like collecting and publishing advertiser identity and verification information. Restrictions are mandated on targeting minors or using sensitive data for targeting without explicit consent. Independent audits must confirm platforms' ad repositories cover qualifying ads comprehensively and accurately.

Risk assessment obligations (Articles 26, 34, 35) VLOPs are subjected to extensive risk assessment requirements to identify and mitigate societal dangers emerging from the misuse or shortcomings of their content moderation systems, algorithmic processes, business models and behavioral designs.

Articles 26, 27, 34 and related recitals impose extensive systemic risk assessment and mitigation requirements on VLOPs to address societal dangers emerging from their services. VLOPs must conduct annual risk assessments evaluating how their content moderation systems, algorithmic processes, business models, and interface designs negatively impact: (i) User security and fundamental rights like non-discrimination, freedom of expression, media freedom (Article 26; Recital 82); (ii) Democratic discourse, electoral processes, public security (Article 26; Recital 82); (iii) Equality issues like gender-based violence or threats to minors (Article 26; Recital 83). Risks must be assessed from illegal content distribution, inauthentic platform use, and lawful but harmful content. Risks from data practices, biased algorithms, and exploitative interface designs should also be evaluated (Articles 27, 34; Recital 84).

Once risks are identified, VLOPs must implement reasonable, targeted and proportionate measures to mitigate them within mandated timeframes (Article 35). Suggested measures include improving content moderation, algorithmic accountability, advertisement and recommender systems, internal processes, third-party auditing, etc. (Article 35). VLOPs must also submit to independent risk assessment audits, share additional data and algorithms for closer regulatory scrutiny upon request, and maintain audit documentation (Articles 26, 37, 40).

Those failing to conduct assessments or mitigate uncovered risks face substantial fines, blacklisting and potential service bans (Articles 73, 74).

Due diligence obligations (Articles 14, 16, 20, 24-27, 34-35) The DSA establishes a graduated set of responsibilities and due diligence duties for online intermediaries based on their role and capacity to cause or mitigate societal risks.

Mere conduit intermediaries have minimal obligations beyond transparency (Article 4). Hosting services must comply with notice-and-action procedures for illegal content (Article 16), inform authorities about detected illegal activities (Article 18), and provide annual transparency reports on content moderation activities (Article 15). Online platforms additionally need to implement complaint and redress mechanisms for content moderation disputes (Articles 20-21). Large online platforms serving significant user bases must also conduct risk assessments and external audits (Articles 26-27). VLOPs shoulder more extensive due diligence requirements, including granting vetted researchers access to platform data to monitor risks (Article 40). Their compliance systems and risk mitigation measures must undergo mandatory independent auditing (Articles 35-37).

VLOP policies should be vetted by external experts to identify inconsistencies with DSA obligations (Recital 105). However, critics argue these duties may still be excessive for smaller platforms. Overall, intermediaries must implement reasonable duties of care, proportional to their resources and risks posed to society based on factors like user base size, reach, impact and capabilities (Recital 40).

Independent auditing mechanisms (Articles 37) To facilitate enhanced accountability of very large intermediaries, the DSA contains extensive provisions concerning independent external audits of platform activities with potential societal impacts. To this purpose intermediaries must implement reasonable duties of care, proportional to their resources and risks posed to society, including a duty to cooperate with audits (Recital 40).

VLOPs must commission annual audits by accredited entities to assess compliance with obligations related to risk management systems (Article 26), content moderation processes (Article 15), algorithmic systems (Article 29), and advertising transparency (Articles 24, 30) (Article 28). Audit summaries containing key methodologies, findings and recommendations must be published, while preserving trade secrets (Article 37). The European Commission and national Digital Service Coordinators can request additional ad-hoc external audits on issues of concern (Article 33).

Targeted internal audits are also imposed focusing specifically on ad transparency (Article 24) and risk assessments (Article 35). Audit findings may inform future regulatory

interventions or enforcement actions if systemic non-compliance is uncovered (Article 35).

However, audits are limited to assessing procedural compliance rather than evaluating substantive fairness, proportionality or effectiveness (Recital 92). The level of genuine independence between powerful platforms and auditors also remains unclear (Article 37).

Delegated act on auditing procedures and methodologies

On October 20, 2023, the European Commission published a delegated act to provide a framework guiding VLOPs and very large online search engines (VLOSEs) in performing audits mandated under Article 37 of the DSA [50]. The act specifies several procedural rules and audit templates:

- Requires using a "*reasonable assurance*" audit standard entailing high confidence in assessing compliance (Annex I).
- Mandates reporting audit risk analysis before conducting procedures (Article 9).
- Requires outlining planned procedures, evidence sources, and sampling methods (Article 10).
- Stipulates providing compliance conclusions for each audited DSA obligation as positive, positive with comments, or negative (Article 8).
- Offers guidance on assessing novel DSA provisions around risk assessments (Article 13), risk mitigation (Article 14), crisis response (Article 15), and auditing itself (Article 16).
- Provides templates for detailed audit reports (Annex I) and audit implementation reports (Annex II).

This aims to ensure rigor, transparency and comparability in inaugural DSA audits. The act delineates cooperation duties between platforms and auditors, and stipulates methodologies for assessing compliance with core DSA provisions.

Quantifying active users for platforms

To be mentioned as a preliminary mechanism to enact the one aforementioned, the DSA requires online platforms and search engines to count average monthly active EU users based on engagement, not just registered users or transactions. Users are considered active if exposed to content on platforms or submitting queries to search engines.

The EU guidance clarifies providers should count all recipients interacting with services, including non-registered users. Mere incidental use should be excluded. Inauthentic users like bots can be discounted, but profiling solely for counting purposes is not permitted. Providers should avoid double counting users across interfaces.

For platforms, recipients on all sides engaging via exposure to or uploading content qualify, including consumers, traders, and advertisers. For search engines, only users submitting queries and seeing results are counted. The published numbers, reflecting methodology and average over 6 months, provide the basis for designating VLOPs and search engines.

While publication is sufficient, providers are encouraged to also notify the numbers and methodology to the Commission and national coordinators to aid oversight. Other data sources may also be used for designations, subject to a right to be heard. This methodology will significantly impact major platforms like Google, Facebook and X which previously only reported registered or active purchasing users. Complying will require recalibrating tracking and counting systems. Ongoing evaluation will help ensure proportionality across diverse entities during DSA phase-in.

Expected platform impacts

These interlocking mechanisms aim to ensure online intermediaries take proactive, risk-based and accountable approaches to mitigate societal harms emerging from their services and design choices.

Overall, the DSA signifies a paradigm shift for online intermediaries, moving away from self-regulation towards comprehensive accountability and public oversight mechanisms. It aims to compel platforms to take a more proactive, transparent and responsible approach to emerging online threats. However, its requirements have sparked concerns about anti-competitive effects and loss of European tech innovators.

The obligations and mechanisms instituted under the DSA are expected to substantially impact major online platforms, which have signed up to previous self-regulatory EU initiatives like the Code of Practice on Disinformation [46] exposed below, but will now face binding statutory duties.

Smaller platforms

Smaller platforms are defined in the DSA as those not reaching the threshold of having an active EU user base comprising over 10% of the EU population.

These smaller intermediaries will need to implement notice-and-takedown procedures for illegal content in line with Article 14, including accessible reporting channels and timely response mechanisms. Under Article 15, they must also inform authorities about detected illegal activities on their services. And all hosting services including smaller platforms are subject to transparency reporting duties on content moderation activities as per Article 13. For qualifying online platforms, additional obligations apply around implementing internal complaint and redress systems for

content moderation disputes under Articles 17 and 18. Those serving significant user bases may also have to conduct risk assessments under Article 26 and can voluntarily opt into transparency around recommender systems as per Article 29.

Critics argue that these compliance burdens may be disproportionate for smaller entities with limited resources [109]. Creating accessible reporting systems, content review teams, and internal redress mechanisms requires significant staffing and investment. Conducting recurring risk assessments also necessitates sizable internal capacity.

There are concerns these obligations may disproportionately impact startups and new innovators aiming to compete with tech incumbents, thereby contributing to anti-competitive effects that favor established giants with sizable compliance budgets. Loss of smaller European platforms is also feared given difficulties in competing with major US players. As such, specific impacts on smaller intermediaries need careful evaluation during initial DSA roll-out. Evidence gathering could help recalibrate obligations to avoid excessive burdens stifling market dynamism and growth. Targeted financial support for compliance may also be warranted.

VLOPs

The most substantial impacts of the DSA will be felt by major online platforms qualifying as VLOPs and VLOSEs. These are intermediaries reaching the threshold of having an active EU user base equivalent to over 10% of the EU's population, set at around 45 million monthly active users.

VLOPs like Google, Facebook, Instagram, X, TikTok and YouTube will have to comply with extensive transparency requirements, risk assessment obligations, independent auditing duties, and enhanced due diligence expectations under the DSA [51].

For transparency, VLOPs will need to provide researchers access to platform data to enable continuous monitoring of risks and research in the public interest. Their annual transparency reports must contain extensive details on content moderation activities, algorithmic systems, and ad targeting.

VLOPs will also have to conduct and publish annual risk assessments evaluating potential harms to electoral processes, public security, civic discourse, user rights, and equality emerging from their services' design, algorithms and policies. Identified risks necessitate implementation of reasonable and proportionate mitigation measures.

These risk assessments and mitigations will be subject to annual independent audits by accredited bodies assessing procedural compliance. Audit summaries containing key findings and recommendations must be publicly released. VLOPs may face significant fines for systematic failures in risk management amounting to up to 6% of global turnover.

Other enhanced due diligence duties for VLOPs include additional obligations related to identity verification for advertisers, as well as vetting of platform policies by external experts to identify inconsistencies with the DSA's requirements.

Overall, major platforms will need large investments in product redesign, content review, risk analysis, transparency systems and oversight mechanisms to meet the DSA's extensive accountability expectations applicable to very large entities. The costs for smaller competitors will be much lower.

Policy context and timeline

The DSA emerged from rising concerns about online harms enabled through the design and use of digital platforms, and the limits of previous self-regulatory initiatives like the Code of Practice on Disinformation. Its obligations aim to compel platforms to take a more responsible approach to content moderation and transparency.

The DSA was proposed by the European Commission in December 2020 alongside the Digital Markets Act [59], as part of a comprehensive approach to regulating digital services and promoting online safety. Following political agreement between EU co-legislators, it entered into force on 16 November 2022.

The DSA introduced a phased rollout of obligations to provide platforms time to adapt. By 17 February 2023, online intermediaries had to report active EU user numbers to determine qualification as VLOPs or VLOSEs.

On 25 April 2023, the Commission formally designated 17 VLOPs and 2 VLOSEs reaching over 45 million monthly EU users [51]. These entities had 4 months to fully comply with extensive DSA obligations around risk assessments, independent auditing, transparency reporting, content moderation, advertising rules, and due diligence.

By 17 February 2024, EU member states had to establish their national Digital Services Coordinators responsible for DSA supervision and enforcement [52]. This was also the deadline for all other online platforms to comply with baseline DSA rules on issues like notice-and-takedown procedures and internal complaint systems. The phased introduction aims to focus initial oversight on very large entities with the most influence, before expanding coverage. However, smaller platforms have warned of significant compliance burdens, and in February 2024 Coordinators figures were yet not fully operational [121, 123].

As a final note, continued assessment during roll-out was a strategic preparation to ensure proportionality across diverse intermediaries. Indeed, even before the DSA's formal roll-out deadlines platforms and private consultancy firms begun internal implementation efforts as revealed by published readiness checklists outlining provisions [173],

alongside the institution of distinct reporting flows for locally illegal content under Article 14's notice-and-action provisions. For instance, YouTube and Pinterest provided feature customized complaint forms and handling by legal teams [116] before phase-in dates.

Appendix 3: Methodological appraisal

Coding DSA critiques categorization in Sect. 3

The categorization and analysis of critiques in Sect. 3 was developed through a comprehensive qualitative coding methodology based on grounded theory principles [97]. The DSA adopted legal text, academic literature containing conceptual analyses and EU policy papers examining provisions were consulted as data sources.

An initial open coding pass was conducted by segmenting texts into coherent parts and inductively assigning descriptive codes to salient critique phrases across sources [154]. Axial coding related these open codes into coherent groupings centered on legal, technical, psychological, behavioral, and ethical dimensions, consulting similar with established framework for examining technology regulation (e.g. multiple streams framework, work domain analysis and others [9, 14, 27, 75, 167]. Finally, focused coding synthesized the evidence into a taxonomic critique structure reflecting the refined categories [154]. Audit trails documented the iterative analysis process.

Initial Data Gathering Three types of texts were compiled as data sources for the initial open coding phase:

- The foundational legal text of the DSA [57] was extensively analyzed to surface limitations and critiques expressed in academic literature regarding specific provisions. All articles and recitals were examined to extract salient concerns.
- Relevant academic papers or grounded blog articles containing conceptual analyses of the DSA were reviewed, including [23, 78, 82, 104, 107, 109, 160, 169, 183]. Critique passages were identified through close reading.
- Policy analyses published by digital rights organizations examining DSA provisions were consulted, e.g. reports and articles from [42] and Epicenter Network [11, 12]. These provided additional critical perspectives from impacted stakeholders.

This combination of primary legal text, secondary scholarly analyses, and stakeholder policy papers enabled gathering critiques from diverse expert sources grounded in specific DSA provisions.

Open Coding The compiled source texts were segmented into coherent argument components focused on a specific theme or provision. For example, a passage discussing limitations of the notice-and-action approach was one segment. Each segment was read closely, and salient critique phrases were highlighted and assigned a descriptive code summarizing the critique. For instance, the phrase "excessive platform discretion over definitions of illegal content" was coded as "definitional ambiguity". The goal was inductive extraction of key DSA critiques through line-by-line open coding of multiple sources.

Axial Coding The open critique codes were analyzed for connections based on the dimension or theme of the critique. Related codes were grouped into coherent conceptual categories aligned with established frameworks for examining technology regulation as aforementioned. This structured the critiques along five major categories: legal, technical, psychological, behavioral and ethical limitations. For example, the code definitional "ambiguity" was grouped under legal critiques while "oversight gap" was categorized as a technical critique.

Focused Coding Finally, focused coding integrated evidence from the open codes into structured critiques synthesized within each taxonomic category. For example, multiple open codes related to definitional issues were compiled into a coherent legal critique. Audit trails using code memos documented the iterative analysis process. The resulting taxonomy reflects a comprehensive landscape of multidisciplinary critiques compiled through systematic coding of multiple data sources.

The iteration, coded in Python, is as follows. We began by extracting textual information from each .pdf document using the PyPDF2 package. To perform topic modeling on our documents of interest, we employed BERTopic, a highly customizable technique. Notably, BERTopic utilizes transformers based on real text, which means that preprocessing tasks such as stopword removal, lemmatization, and tokenization were handled internally. To ensure comprehensive results, we conducted both multi-document and single-document topic extraction. The latter approach allowed us to preserve the source of each document. Finally, we generated a Similarity Matrix of clustered topics to evaluate the consistency of documents, identify outliers, and establish a baseline for manual category composition. This method was adopted to derive categories for our taxonomies, and we provide the code used for these experiments.

Methodology for coding research goals, research methods, and stakeholders in Sect. 4

The initial methodology involved a general qualitative interpretative process. The overview text was read carefully to extract research goals, approaches, and stakeholders. These were compiled into a table to visualize connections, which were then verified against the original text. Interpretative

rationales for each connection were developed based on the textual context. The revised methodology employed a systematic qualitative coding process to add rigor and specificity. The text was divided into coherent segments by research approach, and two iterative coding passes were performed.

The first was an open-ended, inductive coding to surface salient concepts, followed by a targeted pass to extract specific goals, stakeholders, and connections. An audit trail was maintained to document the coding process and mapping revisions. The mapping of research approaches to goals and involved stakeholders was conducted through a comprehensive interpretative process, as follows:

1. The overview text was carefully read to identify all stated research goals, approaches, and stakeholders relevant to investigating the DSA's impacts.
2. For each research approach, the explicit goals linked to that technique in the text were extracted and recorded. For example, the overview states computational analysis can be used to "audit warning labels for misinformation"—this was documented as a goal.
3. Stakeholders described in connection to each approach were also extracted. For instance, the text discusses network analysis to "map collaboration networks" involving platforms and authorities. So these were recorded as stakeholders for computational analysis.
4. Synonymous goals were consolidated where redundancy was detected. For example, "assessing policy/standards changes" and "analyzing online discourse shifts" were merged under discourse analysis.
5. The final synthesized goals and stakeholders for each approach were compiled in a table to visualize the mapping. Connections were verified against the original text.
6. An interpretative rationale was developed for each connection based on the contextual links described in the overview text. For example, comparative case studies were tied to platforms and researchers as stakeholders because the text states researchers can use case studies to "compare platform responses across events."
7. The interpretative rationales were further strengthened through iterative rounds of qualitative coding to surface latent meanings in the description text. This validated the contextual fidelity of the mapping.
8. The finalized mapping illustrates the synthesized connections between techniques, goals, and stakeholders in a clear, rigorous and reproducible way grounded in the source text. The in-depth methodology upholds academic standards of integrity.

Acknowledgements Contribution from the ITN project NL4XAI and HYBRIDS which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860621 and 101073351. Also, supported by the Galician Ministry of Culture, Education, Professional Training and University (grants ED431G2019/04 and ED431C2022/19 co-funded by ERDF/FEDER program). This work reflects only the authors' views and the European Research Executive Agency (REA) and European Commission are not responsible for any use that may be made of the information it contains. The authors wish to personally acknowledge the Italian NGO Privacy Network and Matteo Navacci for the feedback provided.

Author Contributions Conceptualization: Luca Nannini. Methodology: Luca Nannini. Formal analysis and investigation: Luca Nannini, Eleonora Bonel, Davide Bassi, Michele Maggini. Writing—original draft preparation: Luca Nannini. Writing—review and editing: Eleonora Bonel, Davide Bassi, Michele Maggini, Funding acquisition: Luca Nannini, Davide Bassi, Michele Maggini. Resources: Luca Nannini, Eleonora Bonel, Davide Bassi, Michele Maggini. Supervision: Luca Nannini. All listed authors have approved the manuscript before submission.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data availability This study involved a qualitative coding analysis of existing public documents, including the EU Digital Services Act legal text, academic literature, and policy papers. The specific documents analyzed are cited throughout the article and listed in the References section. The coding methodologies are detailed in Appendices 3.1 and 3.2. No new datasets were generated during the current study.

Declarations

Conflict of interest The authors declare that they have no competing interests, financial or non-financial, that could be perceived as influencing the work submitted for publication.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Al-Sibai, N., Christian, J.: Buzzfeed is quietly publishing entire AI-generated articles. Futurism (2023). https://futurism.com/buzzfeed-publishing-articles-by-ai.https://futurism.com/buzzfeed-publishing-articles-by-ai. Accessed 15 Mar 2024
2. Albert, J.: Platforms' promises to researchers: first reports missing the baseline. AlgorithmWatch (2023). https://algorithmwatch.org/en/platforms-promises-to-researchers/. Accessed 15 Mar 2024
3. Altay, S., Berriche, M., Heuer, H., et al.: A survey of expert views on misinformation: definitions, determinants, solutions, and future of the field. Harv. Kennedy Sch. Misinf. Rev. **4**(4), 1–34 (2023). https://doi.org/10.37016/mr-2020-119

4. Alturayeif, N., Luqman, H., Ahmed, M.: A systematic review of machine learning techniques for stance detection and its applications. Neural Comput. Appl. **35**(7), 5113–5144 (2023). https://doi.org/10.1007/s00521-023-08285-7
5. Arcos, R., Gertrudix, M., Arribas, C., et al.: Responses to digital disinformation as part of hybrid threats: a systematic review on the effects of disinformation and the effectiveness of fact-checking/debunking. Open Res. Europe **2**(8), 8 (2022). https://doi.org/10.12688/openreseurope.14088.1
6. Arora, S.D., Singh, G.P., Chakraborty, A., et al.: Polarization and social media: a systematic review and research agenda. Technol. Forecast. Soc. Change **183**, 121942 (2022). https://doi.org/10.1016/j.techfore.2022.121942. https://www.sciencedirect.com/science/article/pii/S0040162522004632
7. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings (2015). arXiv:1409.0473
8. Balkin, J.M.: Free speech in the algorithmic society: big data, private governance, and new school speech regulation. UCDL Rev. **51**, 1149 (2017). https://doi.org/10.2139/ssrn.3038939
9. Bamberger, K.A.: Technologies of compliance: risk and regulation in a digital age. Texas Law Rev. **88**, 669–729 (2010). https://ssrn.com/abstract=1463727, UC Berkeley Public Law Research Paper No. 1463727
10. Barata, J., Budzinski, O., Cole, M., et al.: Unravelling the Digital Services Act package. Technical report. European Audiovisual Observatory, Strasbourg (2021). https://ora.ox.ac.uk/objects/uuid:cfd633e1-57b2-40d1-9540-b67d2e7f4cb1/files/rpc289k03t, editor: Maja Cappello
11. Barczentewicz, M.: The DSA and small and medium enterprises as users of online services. EPICENTER (2021). https://www.epicenternetwork.eu/research/briefings/the-dsa-and-small-and-medium-enterprises-as-users-of-online-services/. Accessed 15 Mar 2024
12. Barczentewicz, M.: The New European Digital Services Act: risky for consumers and innovation. EPICENTER (2021). https://www.epicenternetwork.eu/research/briefings/the-new-european-digital-services-act-risky-for-consumers-and-innovation/. Accessed 15 Mar 2024
13. Barni, M., Fang, Y., Liu, Y., et al.: Combating misinformation/disinformation in online social media: a multidisciplinary view. APSIPA Trans. Signal Inf. Process. **11**(2), e42 (2022). https://doi.org/10.1561/116.00000127
14. Bashir, M.F., Ma, B., Bashir, M.A., Radulescu, M. and Shahzad, U.: Investigating the role of environmental taxes and regulations for renewable energy consumption: evidence from developed economies. Econ. Res. Ekonomska Istraživanja **35**(1), 1262–1284 (2022). https://doi.org/10.1080/1331677X.2021.1962383
15. Beaudoin, C.E.: Mass media use, neighborliness, and social support: assessing causal links with panel data. Commun. Res. **34**(6), 637–664 (2007). https://doi.org/10.1177/0093650207307902
16. Bender, E.M., Koller, A.: Climbing towards NLU: on meaning, form, and understanding in the age of data. In: Jurafsky, D., Chai, J., Schluter, N., et al. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020. Association for Computational Linguistics, pp. 5185–5198 (2020). https://doi.org/10.18653/v1/2020.acl-main.463
17. Benkler, Y., Faris, R., Roberts, H.: Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics. Oxford University Press, Oxford (2018). https://doi.org/10.1093/oso/9780190923624.001.0001. https://academic.oup.com/book/26406/book-pdf/49828327/9780190923648_web.pdf
18. Blondé, J., Easterbrook, M.J., Harris, P.R., et al.: Taking advantage of multiple identities to reduce defensiveness to personally threatening health messages. Appl. Psychol.: Health Well-Being **14**(3), 862–880 (2022). https://doi.org/10.1111/aphw.12355
19. Bovet, A., Makse, H.A. Influence of fake news in Twitter during the 2016 US presidential election. Nat Commun **10**, 7 (2019). https://doi.org/10.1038/s41467-018-07761-2
20. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al.: Language models are few-shot learners. Adv. Neural Inf. Process. Syst. **33**, 1877–1901 (2020). https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
21. Bruns, A.: After the 'apicalypse': social media platforms and their fight against critical scholarly research. Inf. Commun. Soc. **22**(11), 1544–1566 (2019)
22. Buijs, D.: The Digital Services Act and the implications for news media and journalistic content (Part 1) (2021). https://dsa-observatory.eu/2022/09/29/digital-services-act-implications-for-news-media-journalistic-content-part-1/. Accessed 15 Mar 2024
23. Buiten, M.C.: The digital services act from intermediary liability to platform regulation. J. Intellect. Prop. Inf. Technol. Electron. Commer. Law **12**, 361 (2021). https://doi.org/10.2139/ssrn.3876328
24. Bundtzen, S.: Suggested for You: Understanding How Algorithmic Ranking Practices Affect Online Discourses and Assessing Proposed Alternatives. Technical report. Institute for Strategic Dialogue (2022). https://www.isdglobal.org/isd-publications/suggested-for-you-understanding-how-algorithmic-ranking-practices-affect-online-discourses-and-assessing-proposed-alternatives/. Accessed 10 Jan 2024
25. Burgess, M.: Openai's custom chatbots are leaking their secrets. Wired (2023). https://www.wired.com/story/openai-custom-chatbots-gpts-prompt-injection-attacks/. Accessed 15 Mar 2024
26. Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., Bonnefon, J.-F., et al. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. arXiv preprintarXiv:2401.05377 (2023)
27. Carden, T., Goode, N., Read, G.J., et al.: Sociotechnical systems as a framework for regulatory system design and evaluation: using work domain analysis to examine a new regulatory system. Appl. Ergon. **80**, 272–280 (2019). https://doi.org/10.1016/j.apergo.2017.02.019
28. Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T.L., Bucknall, B., Haupt, A., et al. Black-Box Access is Insufficient for Rigorous AI Audits. arXiv preprint arXiv:2401.14446 (2024)
29. Cauffman, C., Goanta, C.: A new order: the Digital Services Act and consumer protection. Eur. J. Risk Regul. **12**(4), 758–774 (2021). https://doi.org/10.1017/err.2021.8
30. Chase, P.H.: The EU Code of Practice on Disinformation: The Difficulty of Regulating a Nebulous Problem. Technical report. Transatlantic Working Group on Content Moderation Online and Freedom of Expression (2019). https://www.ivir.nl/publicaties/download/Dispute_Resolution_Content_Moderation_Final.pdf. Accessed 15 Mar 2024
31. Chen, A.K.: Compelled speech and the regulatory state. Ind. Law J. **97**, 881 (2022). https://www.repository.law.indiana.edu/ilj/vol97/iss3/3/. Accessed 15 Mar 2024
32. Child, R., Gray, S., Radford, A. and Sutskever, I. Generating long sequences with sparse transformers. arXiv preprint (2019). arXiv:1904.10509
33. Choi, J.A., Lim, K.: Identifying machine learning techniques for classification of target advertising. ICT Express **6**(3), 175–180 (2020). https://doi.org/10.1016/j.icte.2020.04.012. https://www.sciencedirect.com/science/article/pii/S2405959520301090

34. Colliver, C.: Cracking the Code: An Evaluation of the EU Code of Practice on Disinformation (2020). https://www.isdglobal.org/wp-content/uploads/2020/06/isd_Cracking-the-Code.pdf. Accessed 15 Mar 2024
35. Commission E, Directorate-General for Communications Networks C, Technology: A multi-dimensional approach to disinformation—report of the independent High level Group on fake news and online disinformation. Publications Office (2018). https://doi.org/10.2759/739290
36. Common, M.F.: Beyond the usual suspects: a taxonomy of social media regulations in countries with human rights issues. Int. Rev. Law Comput. Technol. **37**(1), 1–28 (2023). https://doi.org/10.1080/13600869.2022.2043093
37. Council of the European Union: 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—analysis of the final compromise text with a view to agreement (2024). https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf. Accessed 15 Mar 2024
38. Crawford, K., Gillespie, T.: What is a flag for? Social media reporting tools and the vocabulary of complaint. New Media Soc. 18:410–428 (2016). https://api.semanticscholar.org/CorpusID:35338454
39. Devlin, J., Chang, M., Lee, K., et al.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, vol. 1 (Long and Short Papers). Association for Computational Linguistics, pp. 4171–4186 (2019). https://doi.org/10.18653/V1/N19-1423
40. Divinsky, Y.: OWASP top 10 LLM risks—what we learned. Vulcan Cyber (2023). https://vulcan.io/blog/owasp-top-10-llm-risks-what-we-learned/. Accessed 15 Mar 2024
41. Ecker, U.K., Lewandowsky, S., Cook, J., et al.: The psychological drivers of misinformation belief and its resistance to correction. Nat. Rev. Psychol. **1**(1), 13–29 (2022). https://doi.org/10.1038/s44159-021-00006-y
42. EDRi: Digital Services Act Archives. (2023)https://edri.org/policy-files/dsa/. Accessed 15 Mar 2024
43. Erhardt, K., Pentland, A.S.: Disambiguating disinformation: extending beyond the veracity of online content. ICWSM (2021). https://doi.org/10.36190/2021.51
44. EU T: 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). https://disinfocode.eu/
45. 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). https://eur-lex.europa.eu/eli/reg/2016/679/oj. Accessed 15 Mar 2024
46. European Commission: Communication from the commission to the European Parliament, the council, the European Economic and Social Committee and the committee of the regions—tackling online disinformation: a European approach (2018). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236. Accessed 15 Mar 2024
47. European Commission: Guidance for strengthening the code of practice on disinformation (2021). https://digital-strategy.ec.europa.eu/en/library/guidance-strengthening-code-practice-disinformation. Accessed 15 Mar 2024
48. European Commission: Assessment of the code of practice on disinformation: achievements and areas for further improvement (2022). https://digital-strategy.ec.europa.eu/en/library/assessment-code-practice-disinformation-achievements-and-areas-further-improvement. Accessed 15 Mar 2024
49. European Commission: The strengthened code of practice on disinformation 2022 (2022). https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation. Accessed 15 Mar 2024
50. European Commission: Delegated regulation on independent audits under the digital services act (2023). https://digital-strategy.ec.europa.eu/en/library/delegated-regulation-independent-audits-under-digital-services-act. Accessed 15 Mar 2024
51. European Commission: DSA: very large online platforms and search engines. European Commission—Digital Strategy Portal (2023). https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops. Accessed 15 Mar 2024
52. European Commission: The Digital Services Act package (2023). https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package. Accessed 15 Mar 2024
53. European Commission, European Parliament, European Council, et al.: Action plan against disinformation. Technical Report. JOIN(2018) 36 final, European Commission (2018). https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:52018JC0036
54. European Federation of Journalists (EFJ): EU Digital Services Act: journalistic content must be protected from interference by online platforms (2021). https://europeanjournalists.org/blog/2021/12/03/eu-digital-service-act-european-citizens-need-a-stronger-dsa/. Accessed 15 Mar 2024
55. European Parliament: New EU rules needed to make digital platforms less addictive (2023). https://www.europarl.europa.eu/news/en/press-room/20231023IPR08161/new-eu-rules-needed-to-make-digital-platforms-less-addictive. Accessed 15 Mar 2024
56. European Parliament, Council: Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce') (2000). http://data.europa.eu/eli/dir/2000/31/oj. Accessed 15 Mar 2024
57. European Parliament, 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) (2020). https://eur-lex.europa.eu/eli/reg/2022/2065/oj. Accessed 15 Mar 2024
58. European Parliament, Council: Proposal for a regulation of the European Parliament and of the council on harmonised rules on fair access to and use of data (Data Act) (2022). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2022:68:FIN. Accessed 15 Mar 2024
59. European Parliament, 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 (Digital Markets Act) (2022). https://eur-lex.europa.eu/EN/legal-content/summary/digital-markets-act.html. Accessed 15 Mar 2024
60. European Parliament, Council: Regulation (EU) 2022/868 of the European Parliament and of the council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) (2022). http://data.europa.eu/eli/reg/2022/868/oj. Accessed 15 Mar 2024
61. Fabbri, M.: Self-determination through explanation: an ethical perspective on the implementation of the transparency requirements for recommender systems set by the Digital Services Act of the European Union. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. AIES '23. Association for Computing Machinery, New York, pp. 653–661 (2023). https://doi.org/10.1145/3600211.3604717

62. Farina, M., Yu, X., Lavazza, A.: Ethical considerations and policy interventions concerning the impact of generative AI tools in the economy and in society. AI Ethics (2024). https://doi.org/10.1007/s43681-023-00405-2
63. Fazio, L.K., Rand, D.G., Pennycook, G.: Repetition increases perceived truth equally for plausible and implausible statements. Psychon. Bull. Rev. **26**(5), 1705–1710 (2019). https://doi.org/10.3758/s13423-019-01651-4
64. Fazio, L.K., Pillai, R.M., Patel, D.: The effects of repetition on belief in naturalistic settings. J. Exp. Psychol.: Gen. (2022). https://doi.org/10.1037/xge0001211
65. Forati, A.M., Ghose, R.: Geospatial analysis of misinformation in COVID-19 related tweets. Appl. Geogr. **133**, 102473 (2021). https://doi.org/10.1016/j.apgeog.2021.102473. https://www.sciencedirect.com/science/article/pii/S0143622821000898
66. Fourney, A., Rácz, M.Z., Ranade, G., et al.: Geographic and temporal trends in fake news consumption during the 2016 US presidential election. In: Lim, E., Winslett, M., Sanderson, M., et al. (eds.) Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06–10, 2017. ACM, pp. 2071–2074 (2017). https://doi.org/10.1145/3132847.3133147
67. Fraser, R.: How to talk back: hate speech, misinformation, and the limits of salience. Polit. Philos. Econ. **22**(3), 315–335 (2023). https://doi.org/10.1177/1470594X231167593
68. Gaozhao, D.: Flagging fake news on social media: an experimental study of media consumers' identification of fake news. Gov. Inf. Q. **38**(3), 101591 (2021). https://doi.org/10.1016/j.giq.2021.101591. https://www.sciencedirect.com/science/article/pii/S0740624X21000277
69. Geeng, C., Yee, S., Roesner, F.: Fake news on facebook and twitter: investigating how people (don't) investigate. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20. Association for Computing Machinery, New York, pp. 1–14 (2020). https://doi.org/10.1145/3313831.3376784
70. Ginossar, T., Cruickshank, I.J., Zheleva, E., et al.: Cross-platform spread: vaccine-related content, sources, and conspiracy theories in YouTube videos shared in early Twitter COVID-19 conversations. Hum. Vaccines Immunother. **18**(1), 1–13 (2022). https://doi.org/10.1080/21645515.2021.2003647
71. Gokhberg, L., Meissner, D., Kuzminov, I.: What semantic analysis can tell us about long term trends in the global STI policy agenda. J. Technol. Transf. (2022). https://doi.org/10.1007/s10961-022-09959-5
72. Goldstein, J.A., Sastry, G., Musser, M., et al.: Generative language models and automated influence operations: emerging threats and potential mitigations (2023). CoRR. arXiv:2301.04246
73. Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." Advances in neural information processing systems 27 (2014). https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf
74. Gorwa, R., Binns, R., Katzenbach, C.: Algorithmic content moderation: technical and political challenges in the automation of platform governance. Big Data Soc. **7**(1), 1–15 (2020). https://doi.org/10.1177/2053951719897945
75. Goyal, N., Howlett, M., Taeihagh, A.: Why and how does the regulation of emerging technologies occur? Explaining the adoption of the EU general data protection regulation using the multiple streams framework. Regul. Gov. **15**(4), 1020–1034 (2021). https://doi.org/10.1111/rego.12387
76. Guarino, S., Trino, N., Chessa, A., et al.: Beyond fact-checking: network analysis tools for monitoring disinformation in social media. In: Cherifi, H., Gaito, S., Mendes, J.F., et al. (eds.) Complex Networks and Their Applications VIII, pp. 436–447. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-36687-2_36
77. Guy, B., Alison, G., Elizabeth, O., et al.: Platform problems and regulatory solutions: findings from a comprehensive review of existing studies and investigations. Technical report. UNESCO (2023). https://unesdoc.unesco.org/ark:/48223/pf0000385813
78. Hacker, P., Engel, A., Mauer, M.: Regulating ChatGPT and other large generative AI models. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. FAccT '23. Association for Computing Machinery, New York, pp. 1112–1123 (2023). https://doi.org/10.1145/3593013.3594067
79. Hamed, S.K., Ab Aziz, M.J., Yaakub, M.R.: Fake news detection model on social media by leveraging sentiment analysis of news content and emotion analysis of users' comments. Sensors **23**(4), 1748 (2023). https://doi.org/10.3390/s23041748
80. Harrison, M.: Top google result for "Edward Hopper" an AI-generated fake. Futurism (2023). https://futurism.com/top-google-result-edward-hopper-ai-generated-fake. Accessed 15 Mar 2024
81. Heikkilä, M.: We are hurtling toward a glitchy, spammy, scammy, AI-powered internet (2023). https://www.technologyreview.com/2023/04/04/1070938/we-are-hurtling-toward-a-glitchy-spammy-scammy-ai-powered-internet/?. Accessed 15 Mar 2024
82. Heldt, A.P.: EU Digital Services Act: The White Hope of Intermediary Regulation, pp. 69–84. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-030-95220-4_4
83. Hosseini, M.B., Breaux, T.D., Slavin, R., et al.: Analyzing privacy policies through syntax-driven semantic analysis of information types. Inf. Softw. Technol. **138**, 106608 (2021). https://doi.org/10.1016/j.infsof.2021.106608. https://www.sciencedirect.com/science/article/pii/S0950584921000859
84. Huang, K.: Why Pope Francis is the star of A.I.-generated photos. The New York Times (2023). https://www.nytimes.com/2023/04/08/technology/ai-photos-pope-francis.html. Accessed 15 Mar 2024
85. Humprecht, E., Esser, F., Van Aelst, P., et al.: The sharing of disinformation in cross-national comparison: analyzing patterns of resilience. Inf. Commun. Soc. **26**(7), 1342–1362 (2023). https://doi.org/10.1080/1369118X.2021.2006744
86. Hunt, K., Wang, B., Zhuang, J.: Misinformation debunking and cross-platform information sharing through twitter during Hurricanes Harvey and Irma: a case study on shelters and ID checks. Nat. Hazards **103**, 861–883 (2020). https://doi.org/10.1007/s11069-020-04016-6
87. Hussein, E., Juneja, P., Mitra, T.: Measuring misinformation in video search platforms: an audit study on YouTube. In: Proceedings of the ACM Human Computer Interaction, vol. 4(CSCW), pp. 048:1–048:27 (2020). https://doi.org/10.1145/3392854
88. Hussein, E.A., Eldardiry, H.: Investigating misinformation in online marketplaces: an audit study on amazon (2020). arXiv:2009.12468. https://api.semanticscholar.org/CorpusID:221971175
89. Iwendi, C., Mohan, S., Khan, S., et al.: COVID-19 fake news sentiment analysis. Comput. Electr. Eng. **101**, 107967 (2022). https://api.semanticscholar.org/CorpusID:248301005
90. Jaursch, J.: Here is why digital services coordinators should establish strong research and data units—DSA observatory (2023). https://dsa-observatory.eu/2023/03/10/here-is-why-digital-services-coordinators-should-establish-strong-research-and-data-units/. Accessed 15 Mar 2024
91. Jhaver, S., Boylston, C., Yang, D., et al.: Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. In: Proceedings of the ACM Human Computer Interaction, vol.

5(CSCW2), pp. 381:1–381:30 (2021). https://doi.org/10.1145/3479525
92. Ji, J., Zhu, Y., Chao, N.: A comparison of misinformation feature effectiveness across issues and time on Chinese social media. Inf. Process. Manag. **60**(2), 103210 (2023). https://doi.org/10.1016/J.IPM.2022.103210
93. Jiang, C., Yu, Y., Zhang, X.: Modelling and analysis of misinformation diffusion based on the double intervention mechanism. J. Inf. Sci. (2023). https://doi.org/10.1177/01655515231182076
94. Kattenbeck, M., Elsweiler, D.: Understanding credibility judgements for web search snippets. Aslib J. Inf. Manag. **71**(3), 368–391 (2019). https://doi.org/10.1108/AJIM-07-2018-0181
95. Kaye, B.: Australian mayor readies world's first defamation lawsuit over ChatGPT. Reuters (2023). https://www.reuters.com/technology/australian-mayor-readies-worlds-first-defamation-lawsuit-over-chatgpt-content-2023-04-05/. Accessed 15 Mar 2024
96. Kaye, D.: Speech police: the global struggle to govern the Internet. Columbia Global Reports (2019). http://www.jstor.org/stable/j.ctv1fx4h8v. Accessed 15 Mar 2024
97. Khan, S.N.: Qualitative research method: grounded theory. Int. J. Bus. Manag. **9**(11), 224–233 (2014). https://doi.org/10.5539/ijbm.v9n11p224
98. Kiddle, R., Törnberg, P., Trilling, D.: Network toxicity analysis: an information-theoretic approach to studying the social dynamics of online toxicity. J. Comput. Soc. Sci. (2024). https://doi.org/10.1007/s42001-023-00239-2
99. King, K.K., Wang, B.: Diffusion of real versus misinformation during a crisis event: a big data-driven approach. Int. J. Inf. Manag. **71**, 102390 (2023). https://doi.org/10.1016/j.ijinfomgt.2021.102390. https://www.sciencedirect.com/science/article/pii/S0268401221000839
100. Kirchner, J., Reuter, C.: Countering fake news: a comparison of possible solutions regarding user acceptance and effectiveness. In: Proceedings of the ACM Human Computer Interaction, vol. 4(CSCW2), pp. 140:1–140:27 (2020). https://doi.org/10.1145/3415211
101. Klinenberg, D.: (0) Does deplatforming work? J. Confl. Resolut. https://doi.org/10.1177/00220027231188909
102. Klonick, K.: The new governors: the people, rules, and processes governing online speech. Harv. Law Rev. **131**, 1598 (2017). https://harvardlawreview.org/print/vol-131/the-new-governors-the-people-rules-and-processes-governing-online-speech/
103. Knibbs, K.: Scammy AI-generated book rewrites are flooding amazon. Wired (2024). https://www.wired.com/story/scammy-ai-generated-books-flooding-amazon/. Accessed 15 Mar 2024
104. Kuczerawy, A.: The Good Samaritan that wasn't: voluntary monitoring under the (draft) Digital Services Act (2021). https://doi.org/10.17176/20210112-181758-0, https://verfassungsblog.de/good-samaritan-dsa/. Accessed 15 Mar 2024
105. Lachney, M., Green, B., Allen, M.C., et al.: Ethnocomputing And Computational Thinking. Routledge, London (2021). https://doi.org/10.4324/9781003102991-7
106. Lahby, M., Aqil, S., Yafooz, W.M.S., et al.: Online fake news detection using machine learning techniques: a systematic mapping study. Stud. Comput. Intell. **1001**, 3–37 (2022). https://doi.org/10.1007/978-3-030-90087-8_1
107. Laux, J., Wachter, S., Mittelstadt, B.: Taming the few: platform regulation, independent audits, and the risks of capture created by the DMA and DSA. Comput. Law Secur. Rev. **43**, 105613 (2021). https://doi.org/10.1016/j.clsr.2021.105613. https://www.sciencedirect.com/science/article/pii/S0267364921000868
108. Leerssen, P.: An end to shadow banning? Transparency rights in the digital services act between content moderation and curation. Comput. Law Secur. Rev. **48**, 105790 (2023). https://doi.org/10.1016/j.clsr.2023.105790. https://www.sciencedirect.com/science/article/pii/S0267364923000018
109. Leiser, M.: Reimagining digital governance: the EU's Digital Services Act and the fight against disinformation. SSRN (2023). https://doi.org/10.2139/ssrn.4427493. https://ssrn.com/abstract=4427493. Accessed 15 Mar 2024
110. Lejano, R.: Frameworks for Policy Analysis: Merging Text and Context. Routledge, London (2013). https://doi.org/10.4324/9780203625422
111. Leventhal, T.: The need to up our game in countering disinformation. Comp. Strategy **42**(2), 173–186 (2023). https://doi.org/10.1080/01495933.2023.2182107
112. Lewis, S.: Qualitative inquiry and research design: choosing among five approaches. Health Promot. Pract. **16**(4), 473–475 (2015). https://doi.org/10.1177/1524839915580941
113. Li, B.Z., Nye, M., Andreas, J.: Implicit representations of meaning in neural language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Online, pp. 1813–1827 (2021). https://doi.org/10.18653/v1/2021.acl-long.143
114. Li, J., Chang, X.: Combating misinformation by sharing the truth: a study on the spread of fact-checks on social media. Inf. Syst. Front. **25**(4), 1479–1493 (2023). https://doi.org/10.1007/S10796-022-10296-Z
115. Lo, V.H., Wei, R.: Modeling the dynamic process and adverse effects of misinformation. In: Miscommunicating the COVID-19 Pandemic. Routledge, London, pp. 196–207 (2023). https://doi.org/10.4324/9781003355984
116. Loney, B.: How VLOPs have addressed online regulations: key learnings (2023). https://www.trustlab.com/post/how-vlops-have-addressed-online-regulations-key-learnings. Accessed 15 Mar 2024
117. Lopatto, E.: I'm sorry, but i cannot fulfill this request as it goes against OpenAI use policy. The Verge (2024). https://www.theverge.com/2024/1/12/24036156/openai-policy-amazon-ai-listings. Accessed 15 Mar 2024
118. Maddox, J., Malson, J.: Guidelines without lines, communities without borders: the marketplace of ideas and digital manifest destiny in social media platform policies. Soc. Media + Soc. **6**(2), 2056305120926622 (2020). https://doi.org/10.1177/2056305120926622. Publisher: SAGE Publications Ltd
119. Marcoux, T., Obadimu, A., Agarwal, N.: Dynamics of Online Toxicity in the Asia-Pacific Region. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12237. LNAI, pp. 80–87 (2020). https://doi.org/10.1007/978-3-030-60470-7_9
120. Marrazzo, F.: Doing research with online platforms: an emerging issue network. In: Handbook of Research on Advanced Research Methodologies for a Digital Society. IGI Global, pp. 65–86 (2022). https://doi.org/10.4018/978-1-7998-8473-6.ch006
121. Marsh, O.: Got complaints? Want data? Digital service coordinators will have your back—or will they? AlgorithmWatch (2024). https://algorithmwatch.org/en/dsa-day-and-platform-risks/. Accessed 15 Mar 2024
122. Meel, P., Vishwakarma, D.K.: Fake news, rumor, information pollution in social media and web: a contemporary survey of state-of-the-arts, challenges and opportunities. Expert Syst. Appl. **153**, 112986 (2020). https://doi.org/10.1016/j.eswa.2019.112986
123. Miller, G.: The Digital Services Act is fully in effect, but many questions remain. Tech Policy Press (2024). https://www.techpolicy.press/the-digital-services-act-in-full-effect-questions-remain/. Accessed 15 Mar 2024
124. Mirza, S., Begum, L., Niu, L., et al.: Tactics, threats & targets: modeling disinformation and its mitigation. In: ISOC Network

and Distributed Systems Security Symposium (NDSS) (2023). https://doi.org/10.14722/ndss.2023.23657
125. Mitts, T., Pisharody, N., Shapiro, J.N.: Removal of anti-vaccine content impacts social media discourse. In: WebSci '22: 14th ACM Web Science Conference 2022, Barcelona, Spain, June 26–29, 2022. ACM, pp. 319–326 (2022). https://doi.org/10.1145/3501247.3531548
126. Morrison, S.: How unbelievably realistic fake images could take over the internet. Vox (2023). https://www.vox.com/technology/2023/3/30/23662292/ai-image-dalle-openai-midjourney-pope-jacket. Accessed 15 Mar 2024
127. Nannini, L.: *Voluminous yet vacuous*? semantic capital in an age of large language models. In: Ganapini, M.B., Loreggia, A., Mattei, N., et al. (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, CEUR Workshop Proceedings, vol. 3547. CEUR-WS.org (2023). https://ceur-ws.org/Vol-3547/paper5.pdf
128. Navigli, R., Conia, S., Ross, B.: Biases in large language models: origins, inventory and discussion. J. Data Inf. Qual. (2023). https://doi.org/10.1145/3597307, Just Accepted
129. Neff, G.: The new digital dark age. Wired (2024). 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. Accessed 15 Mar 2024
130. Nenadic, I., Brogi, E., Bleyer-Simon, K.: Structural indicators to assess effectiveness of the EU's code of practice on disinformation. Working Paper 2023/34, European University Institute (2023). https://hdl.handle.net/1814/75558. Accessed 15 Mar 2024
131. Novak, M.: Viral images of Donald Trump getting arrested are totally fake (for now). Forbes (2023). https://www.forbes.com/sites/mattnovak/2023/03/19/viral-images-of-donald-trump-getting-arrested-are-totally-fake. Accessed 15 Mar 2024
132. OpenAI: GPT-4 technical report (2023). CoRR. arXiv:2303.08774
133. O'Brien, M.: Tech experts are starting to doubt that ChatGPT and A.I. "hallucinations" will ever go away: "this isn't fixable". Fortune (2023). https://fortune.com/2023/08/01/can-ai-chatgpt-hallucinations-be-fixed-experts-doubt-altman-openai. Accessed 15 Mar 2024
134. Papadopoulou, O., Makedas, T., Apostolidis, L., et al.: MeVer NetworkX: network analysis and visualization for tracing disinformation. Future Internet **14**(5) (2022). https://doi.org/10.3390/fi14050147, https://www.mdpi.com/1999-5903/14/5/147
135. Papaevangelou, C.: The role of citizens in platform governance: a case study on public consultations regarding online content regulation in the European Union. Glob. Media China **8**(1), 39–56 (2023). https://doi.org/10.1177/20594364221150142
136. Papaevangelou, C.: "The non-interference principle'': debating online platforms' treatment of editorial content in the EU's digital services act. Eur. J. Commun. (2023). https://doi.org/10.1177/02673231231189036
137. Parvizi, J., Hmielowski, J.D.: Breaking the mold: examining the effectiveness of techniques to reduce motivated reasoning. Atl. J. Commun. (2023). https://doi.org/10.1080/15456870.2023.2224482
138. Pazho, A.D., Noghre, G.A., Purkayastha, A.A., et al.: A comprehensive survey of graph-based deep learning approaches for anomaly detection in complex distributed systems (2022). CoRR. arXiv:2206.04149
139. Pennycook, G., Rand, D.G.: Fighting misinformation on social media using crowdsourced judgments of news source quality. Proc. Natl. Acad. Sci. **116**(7), 2521–2526 (2019). https://doi.org/10.1073/pnas.1806781116
140. Perlich, C., Dalessandro, B., Raeder, T., et al.: Machine learning for targeted display advertising: transfer learning in action. Mach. Learn. **95**(1), 103–127 (2014). https://doi.org/10.1007/s10994-013-5375-2
141. Pohjonen, M.: A comparative approach to social media extreme speech: online hate speech as media commentary. Int. J. Commun. **13**, 3088–3103 (2019). https://ijoc.org/index.php/ijoc/article/view/9110
142. Pollicino, O.: Generative AI and the Rediscovery of the Legitimate Interest Clause. Institute for European Policymaking @ Bocconi University (2023). https://iep.unibocconi.eu/publications/generative-ai-and-rediscovery-legitimate-interest-clause. Accessed 15 Mar 2024
143. Popiel, P., Sang, Y.: Platforms' governance: analyzing digital platforms' policy preferences. Glob. Perspect. (2021). https://doi.org/10.1525/gp.2021.19094
144. Primig, F.: The influence of media trust and normative role expectations on the credibility of fact checkers. Journal. Pract. (2022). https://doi.org/10.1080/17512786.2022.2080102
145. Qayyum, H., Zhao, B.Z.H., Wood, I.D., et al.: A longitudinal study of the top 1% toxic twitter profiles. In: Proceedings of the 15th ACM Web Science Conference 2023, WebSci 2023, Austin, TX, USA, 30 April 2023–1 May 2023. ACM, pp. 292–303 (2023). https://doi.org/10.1145/3578503.3583619
146. Radford, A., Narasimhan, K., Salimans, T., et al.: Improving language understanding by generative pre-training (2018). https://openai.com/research/language-unsupervised. Accessed 15 Mar 2024
147. Rieger, A.: Interactive interventions to mitigate cognitive bias. In: Bellogín, A., Boratto, L., Santos, O.C., et al. (eds.) UMAP '22: 30th ACM Conference on User Modeling, Adaptation and Personalization, Barcelona, Spain, July 4–7, 2022. ACM, pp. 316–320 (2022). https://doi.org/10.1145/3503252.3534362
148. Robertson, R.E., Lazer, D., Wilson, C.: Auditing the personalization and composition of politically-related search engine results pages. In: Champin, P., Gandon, F., Lalmas, M., et al. (eds.) Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23–27, 2018. ACM, pp. 955–965 (2018). https://doi.org/10.1145/3178876.3186143
149. Rogers, R.: The Propagation of Misinformation in Social Media: A Cross-platform Analysis. Amsterdam University Press, Amsterdam (2023). https://library.oapen.org/handle/20.500.12657/61940
150. Romano, S., Kerby, N., Angius, R., et al.: Generative AI and elections: are chatbots a reliable source of information for voters? AIForensics, AlgorithmWatch (2023). https://aiforensics.org/uploads/AIF_AW_Bing_Chat_Elections_Report_ca7200fe8d.pdf. Accessed 15 Mar 2024
151. Ropek, L.: OpenAI's Sora is a giant "f*ck you" to reality (2024). https://gizmodo.com/openais-sora-is-a-giant-f-ck-you-to-reality-1851261587. Accessed 15 Mar 2024
152. Ruan, T., Kong, Q., McBride, S.K., et al.: Cross-platform analysis of public responses to the 2019 Ridgecrest earthquake sequence on Twitter and Reddit. Sci. Rep. **12**(1), 1634 (2022). https://doi.org/10.1038/s41598-022-05359-9
153. Ruiz, C.D.: Disinformation on digital media platforms: a market-shaping approach. New Media Soc. (2023). https://doi.org/10.1177/14614448231207644
154. Saldaña, J.: The Coding Manual for Qualitative Researchers. Sage Publications Ltd, Thousand Oaks (2009)
155. Samed, D.: Adobe stock is flooded with AI generated gay pride content. Twitter (2023). https://twitter.com/DeanSamed/status/1658833605882265602. Accessed 15 Mar 2024

156. Sankaran, V.: ChatGPT cooks up fake sexual harassment scandal, names real law professor as accused. The Independent (2023). https://www.independent.co.uk/tech/chatgpt-sexual-harassment-law-professor-b2315160.html. Accessed 15 Mar 2024
157. Savolainen, L.: The shadow banning controversy: perceived governance and algorithmic folklore. Media Cult. Soc. **44**(6), 1091–1109 (2022). https://doi.org/10.1177/01634437221077174
158. Schade, M.: How chatGPT and our language models are developed. OpenAI (2023). https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed. Accessed 15 Mar 2024
159. Schade, M.: How your data is used to improve model performance. OpenAI (2024). https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance. Accessed 15 Mar 2024
160. Schwemer, S.F.: Digital services act: a reform of the e-commerce directive and much more. Prepared for A Savin, Research Handbook on EU Internet Law (2022). https://dx.doi.org/10.2139/ssrn.4213014. Accessed 15 Mar 2024
161. Selbst, A.D., Boyd, D., Friedler, S.A., et al.: Fairness and abstraction in sociotechnical systems. In: Boyd, D., Morgenstern, J.H. (eds.) Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29–31, 2019. ACM, pp. 59–68 (2019). https://doi.org/10.1145/3287560.3287598
162. Seo, H., Faris, R.: Special section on comparative approaches to mis/disinformation. Int. J. Commun. (19328036) 15 (2021). http://hdl.handle.net/1808/31688
163. Simko, J., Tomlein, M., Pecher, B., et al.: Towards continuous automatic audits of social media adaptive behavior and its role in misinformation spreading. In: Masthoff, J., Herder, E., Tintarev, N., et al. (eds.) Adjunct Publication of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, Utrecht, The Netherlands, June 21–25, 2021. ACM, pp. 411–414 (2021). https://doi.org/10.1145/3450614.3463353
164. Spicer, A.: Beware the "botsh*t": why generative AI is such a real and imminent threat to the way we live. The Guardian (2024). https://www.theguardian.com/commentisfree/2024/jan/03/botshit-generative-ai-imminent-threat-democracy. Accessed 15 Mar 2024
165. Stalla-Bourdillon, S., Kramcsák, P.T.: ChatGPT and lawful bases for training AI: a blended approach? The Digital Constitutionalist (2023).https://digi-con.org/chatgpt-and-lawful-bases-for-training-ai-a-blended-approach/. Accessed 15 Mar 2024
166. Staudacher, N.: What is chatGPT? Commonly asked questions about chatGPT. OpenAI (2022). https://help.openai.com/en/articles/6783457-what-is-chatgpt. Accessed 15 Mar 2024
167. Stewart, R.B.: Regulation, innovation, and administrative law: a conceptual framework. Calif. Law Rev. **69**(5), 1256–1377 (1981). https://www.jstor.org/stable/3480247
168. Strowel, A., De Meyere, J.: The Digital Services Act: transparency as an efficient tool to curb the spread of disinformation on online platforms. JIPITEC **14**(1), 66 (2023). https://www.jipitec.eu/issues/jipitec-14-1-2023/5708/strowel_pdf.pdf
169. Sullivan, D.: Unpacking "systemic risk" under the EU's digital service act (2023). https://techpolicy.press/unpacking-systemic-risk-under-the-eus-digital-service-act/. Accessed 15 Mar 2024
170. Supervisor, E.D.P.: Opinion 3/2018. EDPS opinion on online manipulation and personal data. Technical report. EU Publications Office (2018). https://doi.org/10.2760/709177
171. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Ghahramani, Z., Welling, M., Cortes, C., et al. (eds.) Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8–13 2014, Montreal, Quebec, Canada, pp. 3104–3112 (2014). https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html
172. Tomlein, M., Pecher, B., Simko, J., et al.: An audit of misinformation filter bubbles on YouTube: bubble bursting and recent behavior changes. In: Pampín, H.J.C., Larson, M.A., Willemsen, M.C., et al. (eds.) RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021–1 October 2021. ACM, pp. 1–11 (2021). https://doi.org/10.1145/3460231.3474241
173. TrustLab: Digital services act checklist: Trustlab (2023). https://www.trustlab.com/trustlab-resources/digital-services-act-checklist. Accessed 15 Mar 2024
174. Tuparova, E., Tagarev, A., Tulechki, N., et al.: Analyzing the evolution of disinformation content on Facebook—a pilot study. In: Eskenazi, A., Totkov, G., Koychev, I. (eds.) Proceedings of the 15th International Conference Education and Research in the Information Society, Plovdiv, Bulgaria, October 13–14, 2022, CEUR Workshop Proceedings, vol. 3372. CEUR-WS.org, pp. 41–49 (2022). https://ceur-ws.org/Vol-3372/paper05.pdf
175. Urman, A., Makhortykh, M.: How transparent are transparency reports? Comparative analysis of transparency reporting across online platforms. Telecommun. Policy **47**(3), 102477 (2023). https://doi.org/10.1016/j.telpol.2022.102477. https://www.sciencedirect.com/science/article/pii/S0308596122001793
176. Valenzuela, S., Halpern, D., Araneda, F.: A downward spiral? A panel study of misinformation and media trust in Chile. Int. J. Press/Polit. **27**(2), 353–373 (2022). https://doi.org/10.1177/19401612211025238
177. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., et al. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA, pp. 5998–6008 (2017). https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
178. Venkatesh, V., Brown, S.A., Sullivan, Y.W.: Guidelines for conducting mixed-methods research: an extension and illustration. J. Assoc. Inf. Syst. **17**(7), 2 (2016). https://doi.org/10.17705/1jais.00433
179. Verma, P., Oremus, W.: What happens when chatGPT lies about real people? The Washington Post (2023). https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/. Accessed 15 Mar 2024
180. Vishwakarma, D.K., Varshney, D., Yadav, A.: Detection and veracity analysis of fake news via scrapping and authenticating the web search. Cogn. Syst. Res. **58**, 217–229 (2019). https://doi.org/10.1016/j.cogsys.2019.07.004. https://www.sciencedirect.com/science/article/pii/S1389041719301020
181. Volokh, E.: The law of compelled speech. Tex. Law Rev. **97**, 355 (2018). https://texaslawreview.org/the-law-of-compelled-speech/
182. Wardle, C., Derakhshan, H.: Information disorder: toward an interdisciplinary framework for research and policy making. Technical Report. DGI(2017)09, Council of Europe (2017). https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c, with research support from Anne Burns and Nic Dias. Published by the Council of Europe. Photos Council of Europe, October, 2017
183. Wilman, F.: The digital services act (DSA)—an overview. Available at SSRN **4304586** (2022). https://doi.org/10.2139/ssrn.4304586. https://ssrn.com/abstract=4304586

184. Xarhoulacos, C., Anagnostopoulou, A., Stergiopoulos, G., et al.: Misinformation vs. situational awareness: the art of deception and the need for cross-domain detection. Sensors **21**(16), 5496 (2021). https://doi.org/10.3390/S21165496
185. Young, D.G., Jamieson, K.H., Poulsen, S., et al.: Fact-checking effectiveness as a function of format and tone: evaluating factcheck.org and flackcheck.org. Journal. Mass Commun. Q. **95**(1), 49–75 (2018). https://doi.org/10.1177/1077699017710453
186. Zakrzewski, C.: The FTC is investigating whether chatGPT harms consumers. The Washington Post (2023). https://www.washingtonpost.com/technology/2023/07/13/ftc-openai-chatgpt-sam-altman-lina-khan/. Accessed 15 Mar 2024
187. Zawadzki, T., Walecki, T., Świeboda, H., et al.: Introduction to methods of modelling information wars as a 21st century threat. Eur. Res. Stud. J. **XXIII**, 1011–1026 (2020). https://doi.org/10.35808/ersj/1922
188. Zhang, H., Goodfellow, I.J., Metaxas, D.N., et al.: Self-attention generative adversarial networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA, Proceedings of Machine Learning Research, vol. 97. PMLR, pp. 7354–7363 (2019). http://proceedings.mlr.press/v97/zhang19d.html

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.