






CORE ANALYSIS

The Digital Services Act and the psychology of social media content reporting: drawing legal inferences from a behavioural experiment on notice-and-action mechanisms

Pietro Ortolani¹ , Sarah Vahed² , Catalina Goanta³ and Alan G. Sanfey^{2,4} 

¹Faculty of Law, Radboud University, Nijmegen, The Netherlands, ²Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands, ³Molengraaff Institute for Private Law, Utrecht University School of Law, Utrecht, The Netherlands and ⁴Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands

Corresponding author: Pietro Ortolani; Email: pietro.ortolani@ru.nl

(Received 9 September 2024; revised 9 May 2025; accepted 9 July 2025)

Abstract

The Digital Services Act (DSA) is a critical piece of the procedural puzzle of generating a safe social media environment. This article highlights how the DSA's ambitious provisions regarding content moderation rely on psychological assumptions and inferences about behaviour, yet do so in the absence of extensive empirical behavioural evidence. The difference between how individuals are assumed to behave and evidence on how they actually behave is a crucial element in the complex landscape of effectively moderating content on social media. For this reason, the article sheds light on the value of behavioural research in understanding preferences, incentives and decisions, and the role of such information in interpreting and developing legal provisions. Specifically, through explaining the results of a novel experimental study designed with the DSA in mind, the article offers insights on certain decision-making processes, and how this evidence can steer away from unfounded suppositions to help interpret and apply the DSA against the reality of behaviour.

Keywords: Digital Services Act; Notice and action mechanisms; behavioural science; empirical legal studies; behavioural law

1. Introduction: psychology as both a cornerstone and a blind spot in the DSA

The Digital Services Act¹ (DSA) entered into force on 1 November 2022 and is fully applicable since 17 February 2024.² This new Regulation, which constitutes one of the pillars of the EU Digital Single Market Strategy,³ overhauls the 2000 eCommerce Directive,⁴ pursuing the ambitious policy goal of opening up a pan-European market for digital services while ensuring the effective and consistent enforcement of national and EU law throughout the European digital space. Indeed, European Commission President Ursula von der Leyen presented the DSA as

¹Regulation (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), OJ L 277, 27.10.2022, p 1–102.

²Art 93(2) DSA (with exceptions concerning specific DSA provisions, which apply from 16 November 2022).

³European Commission, 'A Digital Single Market Strategy for Europe', COM(2015) 192 final; C Busch and V Mak, 'Putting the Digital Services Act in Context' 3 (3021) Journal of European Consumer and Market Law 109.

⁴Directive 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, OJ L 178, 17.7.2000, p 1–16.

giving ‘practical effect to the principle that what is illegal offline, should be illegal online. The greater the size, the greater the responsibilities of online platforms’.⁵

At the core of the DSA are updated rules on internet intermediaries’ liability for illegal online content, with additional obligations for very large platforms and a particular eye on these platforms’ role in countering online illegality.⁶ To this end, the Regulation does not set any bright-line rule on the limits of online freedom of expression; the definition of illegal content is largely left to a complex mosaic of other instruments of national and EU law, as section A of this article will discuss in further detail. Instead, the DSA introduces a new set of avenues for complaint and possibilities of redress. For this reason, the DSA has been aptly described as marking a ‘procedural turn’⁷ in the regulation of the European digital space: by creating complaint and dispute resolution procedures, the Regulation aims to enhance online law enforcement (understood broadly), while refraining from imposing new substantive rules on delicate matters such as online free speech and its limits. In doing so, the Regulation pursues a safer online environment, and it allows a broad range of individuals and entities to play an essential role in achieving this goal.

Among the many complaint avenues regulated in the DSA, notice-and-action mechanisms deserve particular attention, as they provide individuals and entities with an important low-threshold means of reporting illegal content. Pursuant to Article 16, social media platforms (along with other hosting service providers) have an obligation to put in place a mechanism through which individuals and entities can file a ‘notice’, reporting the presence of allegedly illegal content, thus prompting the platform to take ‘action’ (eg, removing the content, restricting its visibility, disabling monetization, suspending the account of the user that posted it, or taking any other relevant moderation measure). The DSA conceives of notice-and-action mechanisms as a tool through which illegal content is detected and promptly removed. More specifically, once platforms become aware (as a result of a notice) of the presence of illegal content on their services, they must act expeditiously to remove that content, and they are liable if they fail to do so.⁸

Next to the notice-and-action mechanism required by Article 16, platforms typically offer the possibility to report (or ‘flag’) content which is not illegal, but which may be incompatible with the relevant platform’s terms of service. Indeed, pursuant to Article 14, platforms are (under certain conditions) able to contractually impose restrictions on the use of their service, with regard to different categories of content that, while not unlawful, may still be harmful or otherwise undesirable. These limitations, which often concern ‘lawful but awful’ (LBA) content (eg, lawful forms of aggressive speech, fake news or pornography), must be clearly set forth by the platforms in their terms and conditions.⁹ Platforms are free to offer a notice-and-action mechanism enabling individuals and entities to report LBA content, but (unlike the case of illegal content) they have no legal obligation to do so. Thus, the DSA is based on a dichotomy of legal vs illegal content: the two categories are subject to different legal regimes, entailing different rights for individuals and entities, and different obligations for the platforms. It remains to be seen whether this dichotomy will be useful in practice, as early empirical investigations into the content moderation strategies of Very Large Online Platforms (VLOPs), as visible in the DSA Transparency Database, show that up

⁵European Commission, ‘Digital Services Act: Commission welcomes political agreement on rules ensuring a safe and accountable online environment’, 23 April 2022 <https://ec.europa.eu/commission/presscorner/detail/en/IP_22_2545> accessed 3 May 2023.

⁶C Cauffman and C Goanta, ‘A new order: the digital services act and consumer protection’ 12 (2021) *European Journal of Risk Regulation* 758; M Buiten, ‘The Digital Services Act from Intermediary Liability to Platform Regulation’ 12 (2021) *Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 361; A Turillazzi et al, ‘The Digital Services Act: an analysis of its ethical, legal, and social implications’ 15 (2023) *Law, Innovation and Technology* 83.

⁷Busch and Mak (n 3) 109–14.

⁸Art 6(1)(b) DSA.

⁹Art 14(1) DSA.

to 99.8 per cent of content is moderated as Terms of Service violation, as opposed to 0.2 per cent of content moderated on the basis of DSA illegality.¹⁰

When observing the notice-and-action mechanism required by Article 16 from the point of view of the behaviour of the individuals and entities that have access to it, some interesting considerations come to light. First, even though these individuals and entities may lack legal expertise, they are expected to make their own broad-brush evaluation as to whether a certain piece of content may be illegal (and thus warrant a notice) or not. Furthermore, the DSA assumes that, when these individuals and entities encounter online content that they deem to be illegal, irrespective of its nature or severity, they will consider the possibility to file a notice and alert the relevant platform. This, in turn, entails that people will trust both platforms in general, and notice-and-action mechanisms in particular to handle their report. In this regard, trust broadly refers to the belief that others (including authoritative or organisational bodies) will act in ways that serve in the interest of ourselves and the broader community.¹¹ Within the realm of digital platforms, trust can more specifically be defined as ‘the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party’.¹² Indeed, individuals and entities may not make use of the mechanism regulated in Article 16, if they believe that a notice-and-action mechanism is not reliable and effective as a reaction against potentially illegal content.¹³ Interestingly, to date, there is a paucity of empirical, specifically experimental, evidence to support these psychological assumptions.

To further complicate the picture, on top of the aforementioned implicit assumptions, the DSA occasionally makes explicit reference to the judgements and decision-making processes of people. For instance, the Regulation requires those submitting a notice to enclose a ‘statement confirming the bona fide belief [...] that the information and allegations contained [in the notice] are accurate and complete’.¹⁴ The DSA, thus, expressly touches upon the evaluative judgements of the person submitting a notice: while that individual or entity may lack the legal expertise necessary to ascertain whether the content is indeed illegal, they must at least believe in good faith that the allegations are well-grounded. Along similar lines, Article 23(3)(d) of the DSA requires platforms to take user ‘intention’ into account, if possible, when deciding whether to suspend a user that misused the platform (eg, by repeatedly posting manifestly illegal content or filing manifestly unfounded notices). Thus, assumptions about individuals, including about their decision-making, are interwoven with not only the regulatory architecture, but also the text itself of the DSA.

Despite the importance of the psychological assumptions embedded within the legal mandates and goals of the DSA, to date, there is limited empirical evidence seeking to understand how individuals or entities will indeed approach reporting online content on social media platforms in

¹⁰R Kaushal et al, ‘Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database’ (2024) ACM FAccT (forthcoming).

¹¹In psychological research, trust in organisational bodies refers to the extent to which people think that such authority is ‘concerned about others’ well-being and will act in ways that will serve the interests of these others’, and has in this way been studied as an antecedent of procedural fairness judgements and people’s willingness to co-operate with authorities across a wide range of social situations, as noted in D De Cremer and T Tyler, ‘The Effects of Trust in Authority and Procedural Fairness on Cooperation’ 92(3) (2007) *Journal of Applied Psychology* 639.

¹²T Duong Oesterreich et al, ‘What Drives Individuals’ Trusting Intention in Digital Platforms? An Exploratory Meta-Analysis’ (2024) *Management Review Quarterly* 1.

¹³The available data suggests that, to date, Art 16 notices play a relatively minor role in content moderation on social media platforms. For instance, in the month of March 2025, Meta adopted just 30.447 content moderation measures originating from an Art 16 notice with regard to Facebook. For scale, in the same month, a total of 57.451.746 measures were adopted by Meta with regard to Facebook. Furthermore, of all measures originating from an Art 16 notice, only 1 statement of reason indicated illegality as the legal ground, while the other ones referred to a violation of the platform’s terms of service. A similar picture emerges with regard to other social media platforms: in the same month, TikTok adopted just 16.275 content moderation measures on the basis of Art 16 notices, out of a total of 201.931.930 statements of reason.

¹⁴Art 16(2)(d) DSA.

light of these new procedural requirements.¹⁵ What influences the decision whether to report content or not? What type of content is perceived as warranting such notice? Do individuals feel a responsibility to file notices? Do they interpret content moderation as a means of punishment of the users that posted the content, and would they endorse punishment as an appropriate method of redress for such actions? More generally, do individuals trust content moderation, and social media platforms as a whole? To date, these questions remain largely untested and unanswered. In the absence of any empirical evidence upon which to understand individuals' judgements and decisions, the DSA suffers from three blind spots. First, it is unclear whether the Regulation's assumptions and overall architecture comport with the reality of behaviour. Second, as mentioned above, some provisions of the Regulation expressly refer to individual judgements (eg, 'intention' and 'bona fide belief'), but scarce behavioural data exists to guide platforms and courts in the enforcement of these provisions, and it is difficult to identify and retrieve empirical data across different platforms.¹⁶ Third, in the absence of extensive empirical evidence, it is challenging to assess to what extent certain provisions of the DSA are likely to have a practical impact. Identifying these blind spots will help sketch a future research agenda and assess the need for potential reform, with a view to the upcoming review of the Regulation in 2027.¹⁷

The remainder of this article aims at working towards evidence-informed answers to the questions outlined above. To this end, we conducted an experimental study with individuals using social media, aimed at testing their assumptions and considerations, discussed in section 2 below. Against that background, section 3 draws a number of legal inferences from the empirical evidence, with specific reference to the DSA. To be sure, empirical data does not automatically result in a normative claim, nor do we argue that the results of our study are, in and of themselves, sufficient to justify a reform of the DSA. However, as section 3 will illustrate, looking at the Regulation through the prism of our empirical (and specifically experimental) results does shed an interesting light on several provisions of the DSA, as well as paving the way for more research on the topic. Finally, section 4 summarises the main conclusions, taking stock of areas for future research and charting unanswered questions which require further examination.

2. Experimental evidence

This section provides a concise summary of experimental behavioural research in general, and our study in particular, describing its goals and results for purposes of drawing legal inferences about the DSA.

A. Experimental research in behavioural science

Before moving to the details of our study, it is worth briefly highlighting the role of research on human behaviour, especially when used in the context of understanding social media users as well as

¹⁵Empirical evidence exists concerning users' perceived severity of harmful online content: J Aaron Jiang et al, 'Understanding International Perceptions of the Severity of Harmful Content Online' 16 (2021) *PloS One* e0256762; MK Scheuerman et al, 'A Framework of Severity for Harmful Content Online' 5 (2021) *Proceedings of the ACM on Human-Computer Interaction* 1. In addition, survey-based and qualitative research in media studies and platform governance has addressed perceptions of content moderation and practices surrounding flagging mechanisms, user strategies and mechanisms (eg, S Myers-West, 'Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms' 20 (2018) *New Media & Society* 4366; N Suzor et al, 'What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation' 31 (2019) *International Journal of Communication* 1002; K Crawford and T Gillespie, 'What is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint' 15 (2014) *New Media & Society* 410; R Ma et al, 'How do users experience moderation? A systematic literature review' 7 (2023) *Proceedings of the ACM on Human-Computer Interaction* 1; A Kalch and T Naab, 'Replying, Disliking, Flagging: How Users Engage with Uncivil and Impolite Comments on News Sites' 6(4) (2017) *Studies in Communication and Media* 395). That evidence, however, does not experimentally test the psychological drivers that underlie engagement with notice-and-action mechanisms, which the DSA places at the center of online content moderation.

¹⁶J Van Bavel et al, 'Social Media and Morality' 75 (2024) *Annual Review of Psychology* 311.

¹⁷Art 91(2) DSA.

constructing and interpreting laws such as the DSA. At the outset, it is important to clarify that this discussion does not aim to provide a comprehensive overview of all empirical methodologies employed in research investigating behaviour (or empirical legal studies in general).¹⁸ Such research encompasses a diverse array of qualitative, quantitative and mixed methods, each with distinctive strengths and limitations that make them suitable for addressing specific research questions. For instance, qualitative approaches (which include ethnographic studies, interviews, and thematic analyses) can deliver rich, contextually nuanced insights into people's perceptions and experiences, enabling a nuanced understanding of complex social processes. By contrast, quantitative methods (including survey-based and experimental studies), are particularly suited to systematically test relationships between specific factors underlying behaviour, allowing researchers to identify patterns and make robust generalisations about human judgements and decision-making. Given this wide methodological landscape, and due to the particular objectives of our current research, we focus here on only one quantitative empirical approach, experimental research.

Broadly defined, behavioural science refers to the study of human behaviour.¹⁹ Researchers within this field formulate hypotheses about how people, for example, make judgements, form attitudes and beliefs, respond to incentives or reach decisions, by relying on established scientific theories and using statistical methods to reach conclusions about the accuracy of these predictions. Experimental research is one means used within the field in which individual judgements and choices are captured in a controlled setting.²⁰ By altering a factor (known as the independent variable) in order to see whether it causes a change in another factor (known as the dependent variable), it is thus possible for researchers to examine, for example, whether changing the type of content that somebody sees, affects a person's decision whether to report such content or not. In this way, experimental research has the particular goal of providing a detailed, mechanistic understanding of judgements and decisions, including the specific psychological factors that underlie such behaviour.²¹

As with all research methods, an experimental approach involves certain trade-offs. Achieving a high degree of experimental control needed to describe relationships and draw inferences typically requires researchers constructing scenarios that approximate, but necessarily simplify, real-world situations. This simplification affects what researchers refer to as ecological validity, namely the degree to which experimental findings can directly generalise or transfer to more complex, naturally occurring settings. While perfect replication of real-world complexity within controlled experimental conditions is inherently challenging, it is important to emphasise that this simplification serves a deliberate and valuable purpose: it allows researchers to precisely isolate

¹⁸A comprehensive overview of methods for empirical legal studies, including research investigating human behaviour, is usefully provided in L Epstein and A Martin. *An introduction to empirical legal research* (Oxford University Press 2014) and, more recently, in C Bijleveld, *Research Methods for Empirical Legal Studies: An Introduction* (Eleven 2023).

¹⁹For a more detailed overview of behavioural science particularly as it applies to legal scholarship, including an introduction to concepts such as 'experimental evidence', see A Tor, 'The Methodology of the Behavioral Analysis of Law' 4 (2008) Haifa Law Review 237.

²⁰As previously noted, we do not attempt to provide an exhaustive account of all quantitative methods available. However, we briefly highlight some additional quantitative approaches frequently employed by researchers. These methods include surveys and questionnaires, where researchers systematically pose structured questions to participants and record their responses in order to explore attitudes, beliefs, or behavioural tendencies; as well as longitudinal studies, in which data are collected repeatedly from the same participants over an extended period, enabling researchers to examine changes and developments in behaviours or perceptions over time. Moreover, techniques employed from fields such as neuroscience, such as a functional Magnetic Resonance Imaging (fMRI) which captures changes in blood flow and oxygenation levels in the brain, can also be used to explore brain responses of individuals in order to elucidate more nuanced understandings of people and the factors underlying their behaviours and preferences.

²¹For an additional overview on the application of behavioural science and experimentation for policy, see R Van Bavel et al, 'Applying Behavioural Sciences to EU Policy-Making' (2013) 26033 European Commission JRC Scientific and Policy Reports <<https://op.europa.eu/en/publication-detail/-/publication/bc430f3c-23d0-4441-8671-d5fc803a7cf6/language-en>> accessed 3 June 2024.

and rigorously test specific factors thought to shape human behaviour. Indeed, the primary strength and objective of experimental research lies precisely in its capacity to uncover clear and reliable causal relationships between selected variables of interest. Nonetheless, the inferences drawn from experimental data thus depend heavily on the scenarios created (including the specific factors the researcher has chosen to study) and the characteristics of the sample populations tested, which might not always reflect broader user populations or authentic online environments. Therefore, while experimental studies can provide significant clarity regarding which psychological factors influence behaviour under carefully defined and controlled conditions, the applicability of these findings to real-world contexts should always be interpreted carefully, considering both the experimental setup and the specific characteristics of the participants involved.

Despite these considerations, behavioural experiments provide valuable insights that can complement and extend traditional approaches used in research in law as well as platform governance. Legal scholarship typically provides normative guidance regarding expected behaviours from individuals and platforms. Qualitative studies, as an empirical example, can offer rich descriptive accounts of perceptions and practices related to online moderation. Behavioural experiments contribute uniquely to this discourse by empirically testing the psychological assumptions underlying people's judgements and decisions. Specifically, an experimental approach enables rigorous testing of how users respond to regulatory provisions, rather than relying on theoretical or normative assumptions about their behaviour. In this way, and more simply put, experimental evidence can help provide data on how people behave, as opposed to how they are assumed to behave. In this regard, and as detailed above, the DSA relies on a number of assumptions about behaviour, and thus experimental studies on this topic can provide important information into how people may indeed interact with the Regulation's procedural requirements. Furthermore, this type of empirical evidence can also help lawmakers consider how more effective provisions ought to be designed in upcoming reviews and amendments of the DSA.²² Ultimately, when integrated with normative legal reasoning and other empirical findings, behavioural experiments contribute to a more robust, empirical foundation for understanding people's behaviour and improving regulatory frameworks in digital governance. Thus, our aim is to underscore how these different disciplinary perspectives are complementary, collectively enriching interdisciplinary dialogue and enhancing empirical legal scholarship including as it pertains to Regulations like the DSA.

B. Empirical findings on reporting and punishment on social media

In order to begin an exploration into judgements and decisions concerning the reporting of online content, we conducted an online experimental study with a diverse sample of social media users residing across EU Members States with the primary goal of capturing their preferences regarding content moderation when presented with content that they may ordinarily be exposed to on social media. The full empirical study, together with a detailed discussion of the statistical analyses and results from a psychological perspective, have been published in a separate article; nonetheless, we still provide our methodology and results below.²³

Methodology

The study entailed presenting a sample of social media users with images of varying moral valences, resembling content commonly found on social media. Specifically, we recruited 300 participants, all

²²Other uses of behavioural research include, amongst others, evaluating the impact of laws and policies as well as garnering insights to identify outcomes of, and disparities resulting from legal requirements and rules. See E Zamir and D Teichman, *Behavioral Law and Economics* (Oxford University Press 2018); C Sunstein, *Behavioral Law and Economics* (Cambridge University Press 2000).

²³S Vahed et al, 'Moral Judgment of Objectionable Online Content: Reporting Decisions and Punishment Preferences on Social Media' (2024) 19 PLOS ONE e0300960.

18 years old or older, fluent in the English language, residing in an EU Member State, and active on social media. To simulate online content, we selected 66 images from the Social-Moral Images Database (SMID). The SMID is a database of images rated by 1812 participants on eight dimensions, including the level of ‘Moral Wrongness’ (1=immoral/blameworthy; 5=moral/praiseworthy) depicted in the image. Thus, by using SMID images, we could rely on a pre-existing measure of the type of moral reaction that these images trigger on average. From the SMID, we chose 22 images with low ‘Moral Wrongness’ ratings (meaning that the SMID participants regarded them as morally negative), 22 images with high scores on the same scale, and 22 images falling in the middle of the range. Importantly for the purpose of DSA-related inferences, none of those images *per se* likely constituted illegal content within the definition of Article 3(h) DSA but did encapsulate a broad range imagery which included depictions of bullying against adults and children, acts of terrorism, political extremism, self-harm and animal abuse. The images were presented to the participants as if they had been shared on social media. These images were also accompanied by different levels of support from a content poster, so the participant knew whether the user supposedly sharing the content wished to share their approval of the depicted Act, or did so as means of expressing their disapproval or criticizing such Act.

We randomly divided our participants into two groups, ‘Control’ and ‘Responsibility’. We told participants in the ‘Control’ group that they would rate real social media posts and should respond using the criteria they would personally use when reacting to social media content. These participants were specifically informed that there were no right or wrong answers to the tasks, and were thus not endowed with any specific responsibility to moderate content. By contrast, we told participants in the ‘Responsibility’ group that they would be rating real social media posts, with the role of ‘user content moderators’. We specifically informed this group that they had the responsibility of identifying harmful and/or inappropriate online content. Beyond this instruction, no other incentive to moderate content was provided to these participants, and the participants were not informed that they had been divided into two groups with different names and tasks. This difference specifically sought to identify whether informing someone that they have the responsibility of identifying harmful and/or inappropriate online content impacts their views on objectionability online.

Judgement and decision-making tasks

Our participants completed two tasks developed to tap into two distinct psychological processes that are relevant for content moderation. In a judgement task, participants were asked to indicate which content they believed should not be shared online and thus wished to report. Participants were then given the opportunity (in a punishment task) to assign a punishment to the user that posted that content, by indicating the length of time for which the user who shared the content should be banned from the platform. Importantly, the punishment task does not reflect the current reality of most popular social media platforms, where individuals and entities are not given the possibility to suggest the type and severity of measure to be applied, when a user posts illegal content. However, giving participants this ability is in line with commonly used measures in behavioural research which can help in elucidating psychological preferences and in drawing evidence-based legal inferences, as section 3 below will illustrate.

Hypotheses

Our study was specifically guided by three primary hypotheses aimed at empirically testing the influence of key factors on participants’ moderation judgements and decisions. First, we hypothesised that the moral nature of the content itself would significantly influence participants’ decisions to report and punish online content. Second, we predicted that the intention of the content’s poster (whether the content was shared approvingly or disapprovingly) would be incorporated into moderation decisions, affecting both decisions to report and preferences to

punish. Third, we anticipated that explicitly assigning responsibility to participants, by instructing some participants to act as ‘user content moderators’, would shift reporting and punishment behaviour. Collectively, these hypotheses aimed to experimentally isolate and clarify the distinct roles these factors play in shaping decision-making within the context of online content moderation.

Results

According to the results of the judgement task, the participants’ decision whether to report content or not was influenced not only by the moral connotation of an image (as rated in the SMID), but also by the intention of the user posting that image, and by the fact that they had been given a ‘user content moderator’ role. First, as far as the moral connotation of the image is concerned, our participants reported morally negative images with a significantly higher frequency than morally positive or neutral images.²⁴ Second, with regard to poster intention, there was a statistically significant difference between the (higher) percentage of participants reporting morally negative content that was being endorsed by the poster, and the (lower) rate of reporting of the same content, when the poster criticised it instead.²⁵ Third, with regard to the participants’ perceived responsibility to file a notice, participants in the ‘Responsibility’ group reported content significantly more frequently than participants in the ‘Control’ group.²⁶

In the punishment task, the moral connotation of the image and the intention of its poster significantly predicted the amount of punishment inflicted by the participants. More specifically, the participants assigned longer bans to users that posted morally negative (as opposed to positive²⁷ or neutral)²⁸ images, and to users that endorsed (rather than criticizing) that content.²⁹ There was, instead, no statistically significant relation between the level of punishment assigned by a participant, and the fact that that participant belonged to the ‘Responsibility’ or ‘Control’ groups.³⁰

The study also revealed insights into individual characteristics and opinions regarding social media generally. When asked whether they had encountered content they would consider inappropriate or harmful in their daily social media use, 99 per cent of participants answered in the affirmative.³¹ Furthermore, in response to the question whether the posters of inappropriate or harmful content should be punished, 75 per cent of participants answered positively.³² Such punishment could take the form (among others) of a content moderation measure adopted by a social media platform, to react against harmful or inappropriate online practices. At the same time, however, the same participants also expressed low or undecided levels of trust towards all major social media platforms covered in the study, with the lowest reported trust observed for Facebook (trust=15%, distrust=69%), TikTok (trust=14%, distrust=59%), Instagram (trust=22%, distrust=58%), and Twitter (now X; trust=29%, distrust=42%).³³ Importantly, a

²⁴OR=441.11; 95% CI [271.07,717.81]; $p < .001$.

²⁵OR=1.20; 95% CI [1.08,1.34]; $p < .001$.

²⁶OR=0.70; 95% CI [0.53,0.93]; $p = 0.01$.

²⁷Negative – positive: OR=30.89; 95% CI [29.24,32.62]; $p = < .001$.

²⁸Neutral – positive: OR=1.10; 95% CI [1.03,1.19]; $p = 0.006$.

²⁹OR=1.24; 95% CI [1.21,1.26]; $p = < .001$.

³⁰OR=1.03; 95% CI [0.83,1.29]; $p = 0.78$.

³¹In response to the question: ‘I’ve seen content I would consider inappropriate or harmful on social media’ (measured with 6-point Likert Scale from ‘Never’ to ‘Always’), 99% of participants answered in the affirmative (9% ‘Very Rarely’, 19% ‘Rarely’, 50% ‘Occasionally’, 14% ‘Very Frequently’ and 7% ‘Always’), with 1% of participants answering ‘Never’.

³²In response to the question: ‘I think those that post content I would consider inappropriate or harmful on social media should be punished’ (measured with 5-point Likert Scale from ‘Strongly Disagree’ to ‘Strongly Agree’), 75% of participants answered positively (52% ‘Agree’ and 23% ‘Strongly Agree’), 18% were ‘Undecided’, and 7% answered negatively (6% ‘Disagree’ and 1% ‘Strongly Disagree’).

³³We operationalized ‘trust’ through two distinct questions, ‘I trust this social media platform’ (where participants were explicitly instructed to reflect on the platform itself, including the companies and owners) and ‘I trust other users on this social

correlation was found between trust towards social media platforms, and rate of reporting of content: Mann-Whitney U tests showed that participants who expressed distrust towards social networking platforms reported fewer images in the Judgement task, compared to those who express trust.³⁴

Taken together, these insights highlight the complexity of content moderation decisions and behaviours which can help to better understand users' role in content moderation, advancing understanding of reporting judgements and punishment preferences on social media, and offering insights for social media platforms and regulatory bodies, as detailed more fully in section 3 below.

3. Legal and research inferences

This section will draw some legal inferences (with specific reference to the DSA) from the empirical evidence summarised in the previous section, as well as use that experimental evidence to identify avenues for future research. More specifically, sub-section A will contrast the psychological processes of reporting content against the legal/illegal dichotomy enshrined in the DSA. Subsequently, sub-section B will scrutinise the relevance of the perceived responsibility to report content under the DSA. Sub-section C, in turn, will consider how the DSA's transparency obligations with respect to content moderation measures affect the perception of (and trust in) content moderation. Finally, sub-section D will discuss how people contextualise content moderation in light of the poster's intention, and what consequences this evidence has for the application of the DSA, and for future research on the topic.

A. Testing the legal/illegal dichotomy against individual preferences

As mentioned above, one of the policy aims of the DSA is to ensure that illegal content is promptly removed from social media platforms. To this end, the Regulation allows individuals and entities to 'flag' or 'report' the allegedly illegal content they encounter online through the notice-and-action mechanism of Article 16, discussed above in section 1. The DSA broadly defines illegal content as 'any information that, in itself or in relation to an activity, [...] is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law'.³⁵ To comply with the DSA, many platforms have included a new notice category, often called 'illegal content', alongside the voluntary notice mechanisms that concern lawful content that is potentially incompatible with the relevant platform's terms of service, as illustrated in section 1 above. Unlike existing typologies of content labels facilitating notification, the category of 'illegal content' prompts the flaggers to identify themselves according to their own legal name, while also having to identify and describe the legal basis upon which illegality is invoked.

When looking at content moderation through a behavioural science lens, a complex reality emerges. Our study demonstrates that the decision to report content is influenced by (among other

media platform' (where participants were unstructured to reflect on other users who they commonly perceive use that platform). Each of the questions was associated with a 5-point Likert Scale ranging from 'Strongly Agree' to 'Strongly Disagree' (with 'Trust' being a composite measure of 'Strongly Agree' and 'Agree', and 'Distrust' being a composite measure of 'Strongly Disagree' to 'Disagree'). The percentages reported here are responses only to the question pertaining to trust in the social media platforms.

³⁴Specifically, for YouTube, participants who trust the platform had a median reporting rate of 15.13%, while those who distrust it had a median reporting rate of 13.64% ($U=5550.5$, $p=.007$). Participants who trust Instagram had a median reporting rate of 16.67%, while those who distrust it had a median reporting rate of 13.64% ($U=4523$, $p=.027$). For Twitter, participants who trust the platform had a median reporting rate of 16.67%, while those who distrust it had a median reporting rate of 13.64% ($U=4015.5$, $p=.002$). Finally, participants who trust Facebook had a median reporting rate of 18.18%, while those who distrust it had a median reporting rate of 13.64% ($U=3045$, $p<.001$). For all analyses, the conventional threshold of $p<.05$ has been used for determining statistical significance.

³⁵Art 3(h) DSA.

factors) moral considerations that transcend the legal/illegal dichotomy: our participants reported morally negative images more often than morally positive or neutral ones, although none of those images were likely to constitute illegal content within the definition of the DSA. In other words, individuals and entities may wish to report or flag online content based on moral considerations, even if that content is not illegal. Thus, a tension emerges between the logic of the DSA and the preferences of individuals. On the one hand, the DSA conceives of the notice-and-action mechanism of Article 16 as a way to enable individuals and entities to report illegal content and obtain its swift removal, consistently with the oft-cited rationale of ensuring that ‘what is illegal offline must also be illegal online’.³⁶ On the other hand, in practice, notice-and-action mechanisms are also used to report morally objectionable content, even in cases where that content is not illegal.³⁷ These results are consistent with the fact that most social media users are not lawyers, and they lack the expertise necessary to determine whether a piece of content they perceive as morally negative is also unlawful. Furthermore, our findings resonate with the results of previous studies, which confirm that morality (rather than only legality) shapes behaviour on social media.³⁸ Thus, Article 16’s exclusive focus on illegal content leaves out other forms of morally negative (but legal) content, which individuals or entities may wish to report, but for which the possibility of reporting depends on the platforms’ voluntary initiative. In the absence of a legal obligation to put in place a notice-and-action mechanism for LBA content, there is a significant risk of blind spots in platform affordances: practical experience demonstrates that, in the current reality of content moderation, platforms often fail to offer the possibility to file certain types of complaint with regard to the content they host.³⁹ This, in turn, raises the issue of disenfranchisement; if somebody wishes to report LBA content, but is not given the opportunity to do so, they may be reluctant to file notices in the future, when they encounter harmful content. Importantly, this could potentially have a negative impact on the practical usefulness of notice-and-action mechanisms as a whole, also as a tool for the detection of illegal content. This is particularly relevant, considering how platforms are currently implementing illegal content into their notice-and-action interfaces: as already mentioned, reporters are asked to reveal their legal identity, and to specifically identify the legal grounds of their report (eg, mentioning the specific law the content infringes, as well as the specific jurisdiction where that law is applicable). While this approach can be a way for platforms to outsource the potential labelling of content as illegal, it also raises serious concerns relating to the reporters’ self-efficacy in performing legal qualification tasks, as well as the impact that such a task may have on the reporters’ incentives to report illegal content.

Furthermore, our study demonstrates that the aforementioned risk of disenfranchisement is far from theoretical. Our participants reported having frequently encountered inappropriate or harmful content on social media,⁴⁰ and showed the conviction that the users that post that content should be ‘punished’.⁴¹ In other words, despite laypeople lacking the legal expertise to distinguish between legal and illegal content, they do routinely encounter content that, in their opinion, warrants moderation, and they think that platforms should take adequate measures to that end. Therefore, it can be argued that, if a platform fails to afford the opportunity to even complain about LBA content (which can and does happen, given that notice mechanisms for LBA content depend on the platforms’ voluntary initiatives), trust in content moderation as a whole may be

³⁶European Commission (n 5).

³⁷To further complicate the picture, the available evidence suggests that the perceived harmfulness of online content varies across cultural contexts: Jiang (n 15); Scheuerman (n 15).

³⁸Van Bavel *et al* (n 16).

³⁹C Goanta and P Ortolani, ‘Unpacking content moderation: The rise of social media platforms as online civil courts’ in X Kramer *et al* (eds), *Frontiers in Civil Justice: Privatisation, Monetisation and Digitisation* (Edward Elgar Publishing 2022) 192.

⁴⁰As mentioned above, in response to the question whether that have seen content they would consider inappropriate or harmful on social media, 99% of participants answered positively.

⁴¹As mentioned above, in response to the question whether the participant thinks those that post content they would consider inappropriate or harmful on social media should be punished, 75% of participants answered positively.

negatively affected. Indeed, our study demonstrates that reporting behaviour is affected by the level of trust towards social media platforms: on average, the participants that expressed distrust towards social media platforms had significantly lower reporting rates than those claiming to trust the same platforms.⁴² To make matters more delicate, our participants expressed on average low and/or undecided levels of trust with respect to all social media platforms covered in our study.⁴³ While further research is needed to unpack the underlying components behind people's (lack of) trust in social media platforms, these findings highlight the importance of trust in online intermediaries in subsequent judgements and decisions.

In sum, from this point of view, the DSA is off to an uphill start: while the Regulation enables active participation in reporting illegal content, individuals and entities may not trust platforms enough to engage with their notice-and-action mechanisms, and that trust may be further eroded by the potential absence of a suitable avenue for complaint with respect to lawful content that is perceived as morally negative. As pointed out by Husovec,⁴⁴ the DSA is geared towards digital due process, and aims at offering an articulate set of avenues for complaint and redress. If disenfranchised individuals end up ignoring those avenues, though, the DSA may end up having limited practical impact. Empirical evidence thus suggests that an extension of Article 16 with respect to LBA content may at least partially alleviate this concern, mitigating the sharp legal/illegal dichotomy on which the provision is currently based.

B. Perceived responsibility and its significance under the DSA

Our study provides insights on the psychological incentives underlying a decision whether to report social media content or not. More specifically, the perception of one's responsibility is a predictive factor in his/her reporting behaviour: if somebody perceives reporting as his/her task or obligation, he/she will file notices more often than somebody perceiving no such responsibility.

This finding casts several DSA provisions in an interesting light, and paves the way for further empirical research on the topic. More specifically, under Article 6 DSA, platforms are generally not liable for online content whose presence they are not aware of. However, once somebody files a notice concerning the presence of illegal content, this gives rise (under certain conditions) to the platform's 'actual knowledge or awareness', depriving the platform of its immunity from liability.⁴⁵ Therefore, platforms may potentially have an incentive not to stimulate people's feeling that they have a responsibility to report certain categories of illegal content. This holds particularly true for content that, whilst illegal, may not be seen as particularly problematic by platforms. A good example are 'advertorials', ie, undisclosed forms of advertisement, which are a widespread problem on social media platforms:⁴⁶ while advertorials are prohibited by EU consumer law,⁴⁷ platforms may not have a strong interest to systematically curtail this practice, which ultimately feeds into the business model of social media, blurring the boundaries between commercial and

⁴²These results comport with previous studies, according to which individuals who trust the authorities or institutions that handle reports of objectionable behavior may be more likely to report such behaviour than those who do not trust these entities (A Feddes and K Jonas, 'Associations between Dutch LGBT Hate Crime Experience, Well-Being, Trust in the Police and Future Hate Crime Reporting' 51 (2020) *Social Psychology* 171; N Guzy and H Hirtenlehner, 'Trust in the German Police: Determinants and Consequences for Reporting Behavior', *Trust and Legitimacy in Criminal Justice: European Perspectives* (Springer 2014) 203).

⁴³The study specifically looked into levels of trust in the following 12 popular social media platforms: Facebook, Tiktok, Instagram, Twitter, Snapchat, YouTube, Twitch, 9gag, Reddit, Pinterest, LinkedIn, and Discord.

⁴⁴Martin Husovec, 'Will the DSA Work? On Money and Effort' (*Verfassungsblog*, 9 November 2022) <<https://verfassungsblog.de/dsa-money-effort/>> accessed 3 June 2024.

⁴⁵Art 16(3) DSA.

⁴⁶B Duivenvoorde and C Goanta, 'The Regulation of Digital Advertising under the DSA: A Critical Assessment' 51 (2023) *Computer Law & Security Review* 105870.

⁴⁷Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market, Annex I, n. 11.

non-commercial speech.⁴⁸ From this point of view, manipulating the perceived responsibility to report illegal content could be a strategy for a platform to avoid losing its immunity from liability, with respect to certain forms of illegal content that the platform has no real interest in countering.

In practice, platforms may be able to alter the perceived responsibility to report content through the information they disclose regarding illegal content and the role of filing notices in targeting them, as well as by structuring the interface and user experience in certain ways which impact one's ability to act. For instance, platforms may make the 'report' button more or less visible, and generally build their interface in a way that emphasises or minimises the perception that a notice can easily be filed. There is even a risk that platforms deploy 'dark patterns',⁴⁹ eg, an intentionally misleading interface and user experience aimed at preventing people from developing any sense of responsibility with respect to certain types of illegal content. Far from being theoretical, these risks have been researched and debated in different contexts, such as Germany's NetzDG law.⁵⁰ Importantly, the DSA addresses these risks, and our findings provide insights into how its provisions could be interpreted and operationalised. More specifically, Article 25 of the DSA focuses on the platform's interface design and organization, precisely with the purpose of avoiding the psychological manipulation of users. Under this provision, platforms are prevented from designing, organizing or operating their online interfaces 'in a way that deceives or manipulates the recipients of their service or in a way that otherwise materially distorts or impairs the ability of the recipients of their service to make free and informed decisions'. Given our findings concerning the perceived responsibility to report content, notice-and-action mechanisms could be monitored closely by the European Commission and the Digital Services Coordinators, when assessing whether platforms comply with Article 25. In addition, further empirical research is necessary to test the impact of different interface design choices on the perceived responsibility to file notices and, thus, their reporting behaviour. This empirical research, in turn, could feed into the guidelines issued by the Commission under Article 25(3), concerning, eg, the prominence of the 'report' button, and/or the pop-ups asking for confirmation when somebody initiates a report.

Furthermore, our findings have the potential to inform the practical application of Article 44(1)(a) of the DSA. Pursuant to this provision, the Commission supports and promotes the development and implementation of voluntary standards in respect of the electronic submission of notices. Standardization of notice-and-action mechanisms can have a meaningful impact on the practical use of these instruments: if all major platforms offered everyone the same mechanism, it would become easier for individuals and entities to familiarise themselves with it, and to navigate the different steps required to submit a notice and trigger the platform's 'actual knowledge or awareness' under Article 6. Article 44, however, does not specify the goals that such a standardization process should pursue. The industry, thus, could standardise notice-and-action mechanisms in a way that encourages the perceived responsibility to file notices, but also in a way that discourages such perception. Standardization may fail to facilitate the filing of notices, and actually hinder it, if the standardised notice mechanism prevents people from developing a sense of responsibility, or otherwise infringes on their ability to act on such responsibility. Our results show that, during the process of standardizing notice-and-action mechanisms, the psychological effect of different design choices could be of paramount importance, and further empirical research is necessary to measure to what extent different aspects of the current notice-and-action mechanisms have an impact on reporting behaviour.

⁴⁸C Goanta and S Ranchordás, 'The Regulation of Social Media Influencers: An Introduction' in C Goanta and S Ranchordás (eds), *The Regulation of Social Media Influencers* (Edward Elgar Publishing 2020) 1.

⁴⁹J Luguri and LJ Strahilevitz, 'Shining a Light on Dark Patterns' 13 (2021) *Journal of Legal Analysis* 43; A Mathur et al, 'Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites' 3 (2019) *Proc ACM Hum-Comput Interact* Article 81; C Gray et al, 'The Dark (Patterns) Side of UX Design' (Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems).

⁵⁰B Wagner et al, 'Regulating transparency? Facebook, Twitter and the German Network Enforcement Act' (2020) *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 261–71.

The findings also have implications for further research on the role of so-called ‘trusted flaggers’. Under Article 22 DSA, trusted flaggers are certified entities with ‘particular expertise and competence for the purposes of detecting, identifying and notifying illegal content’;⁵¹ they are independent from platforms, and they must operate ‘diligently, accurately and objectively’⁵² when submitting notices. According to the same provision, platforms have an obligation to give priority to the notices filed by trusted flaggers, and to process and decide upon them without undue delay.⁵³ Therefore, under the DSA, not all notices are ‘born equal’: notices filed by trusted flaggers are deemed to warrant preferential treatment, compared to notices filed by ‘regular’ flaggers. On the one hand, this differentiation is understandable: since Article 16 of the DSA conceives of notice-and-action mechanisms as tools for the detection of illegal content, it makes sense to grant some kind of priority to notices filed by entities that possess the expertise necessary to assess whether a certain piece of content may indeed be illegal. As mentioned above, flaggers lacking such expertise will often be ill-equipped to evaluate whether content with a morally negative connotation is also illegal. On the other hand, however, if trusted flaggers were to gain particular prominence in content moderation, the question arises whether this could diminish the regular flaggers’ perceived responsibility to file notices, for at least three reasons. First, when encountering potentially illegal content, ‘non-trusted’ flaggers may feel like their notice is superfluous, because certified trusted flaggers are likely to detect and report that content. Second, and relatedly, Article 22 may engender the perception that filing a notice requires sectoral expertise, and that the average social media user is not well equipped to make the decision whether to report content or not. Third, non-trusted flaggers may be discouraged by the fact that their notice will not be processed with priority, and they may assume that their own report will not be regarded as important/relevant by the platform. These hypotheses are in line with the results of our study, which indicated that individuals’ likelihood to take action was significantly influenced by whether they perceived the responsibility to do so as their own.⁵⁴ The primary aim in our experimental study was specifically to examine how experimentally manipulated perceptions of responsibility can (and did indeed) affect participants’ judgements and decisions about reporting potentially harmful content. However, more research is needed before drawing definitive conclusions in this respect: particularly, it is necessary to investigate to what extent the individuals and entities mentioned in Article 16 are aware of the existence of trusted flaggers, and of the preferential treatment accorded to them. A future study could answer this question, and subsequently scrutinise whether this awareness (if existent) indeed has an effect on the incentives to report potentially illegal content.

An additional implication of our study concerns the potential misuse of notice-and-action mechanisms. Article 23 DSA concerns the possibility of ‘misuse’ of social media; more specifically, Article 23(1) addresses the case of a user that frequently provides manifestly illegal content, while Article 23(2) focuses (among other cases) on users that frequently file manifestly unfounded notices. In both scenarios, the DSA requires platforms to suspend the user, ‘for a reasonable period of time and after having issued a prior warning’.⁵⁵ Importantly, under article 23(3)(d), platforms must take into account the ‘intention’ of the user submitting the complaint or notice, when deciding whether that user has abused his/her right to report content. It will be interesting to see how platforms (and courts reviewing platforms’ decisions) will evaluate the intention of those who

⁵¹ Art 22(2)(a) DSA.

⁵² Art 22(2)(c) DSA.

⁵³ Art 22(1) DSA.

⁵⁴ J. Darley and B. Latane, ‘Bystander Intervention in Emergencies: Diffusion of Responsibility’ 8 (1968) *Journal of Personality and Social Psychology* 377; R. Yee Man Wong et al, ‘Standing Up or Standing by: Understanding Bystanders’ Proactive Reporting Responses to Social Media Harassment’ 32 (2021) *Information Systems Research* 561; R. Philpot et al, ‘Would I be Helped? Cross-National CCTV Footage Shows that Intervention is the Norm in Public Conflicts’ 75(1) (2020) *American Psychologist* 66.

⁵⁵ Arts 23(1) and (2) DSA.

file unfounded notices. The question is whether the DSA requires platforms to be more lenient towards users that have demonstrably shown to be feeling a responsibility to report, which could also lead to over-reporting.

Our results on the perceived responsibility to report content also raise some unanswered questions concerning the misuse of notice-and-action mechanisms, paving the way for future research on the topic. More specifically, if people lacking legal expertise perceive a responsibility to report content, platforms are likely to receive a higher number of complaints, compared to the situation where people feel no such responsibility. At the same time, if people do not feel any particular responsibility to report content, and/or if they are affected by the bystander effect of, eg, trusted flaggers, the notices that the platforms receive may be fewer, but more careless. Hence, further research is necessary to explore the relationship between the perceived responsibility to report content, and the frequency of unfounded notices.

C. Transparency obligations and the psychological importance of content moderation as a form of punishment

Our study shows that our participants care about ‘punishment’ of users posting content that is perceived as morally negative. This holds equally true for participants who were given a responsibility to report, and participants who were not. These results tie into the ‘action’ aspect of notice-and-action mechanisms, ie, the content moderation measure that a platform adopts upon receiving a well-founded notice. The DSA does provide transparency in some important respects, for instance requiring the disclosure of what measure has been adopted.⁵⁶ However, there are also some potential limitations and shortcomings. Importantly, the person reporting the content does not receive any explanation in cases where the platform decides not to take any action. In that case, the person that filed the report will only receive a communication that no action has been taken.⁵⁷ Once again, this may lead to disenfranchisement, if the complainant sees that the user goes unpunished, and they do not receive any explanation for the platform’s decision to take no action. Interestingly, this risk is already visible in practice: for instance, starting from 2022, the number of cases submitted by users to Meta’s Oversight Board (the external review body that Meta has set up to hear appeals on delicate moderation decisions made by the Company)⁵⁸ progressively dropped from approximately 150,000 per month to less than 100,000 per quarter,⁵⁹ possibly as a result of the growing user awareness that the Board only selects very few cases for review, while the other ones receive no review.

Furthermore, even in cases where a content moderation measure is adopted, the reporter only receives limited information about the platform’s decision.⁶⁰ While Article 17 of the DSA does impose on the platform an obligation to specifically state the reasons for its decision with a good level of detail, this obligation only exists towards the affected user (ie, the poster), and not towards the complainant. This difference in treatment is not unjustified, since the poster is ultimately the one suffering a limitation of his/her ability to make use of the platform’s services. At the same time, however, it is important that the reporter also receive sufficiently detailed information about what type of measure has been adopted. Indeed, if the reporter only receives an unreasoned communication of the platform’s decision, without sufficient information as to what level of punishment or type of measure has been applied and why, they may end up losing trust in notice-

⁵⁶Art 17 DSA. On the disclosure of the fact that a moderation measure has been adopted see Paddy Leerssen, ‘An End to Shadow Banning? Transparency Rights in the Digital Services Act Between Content Moderation and Curation’ 48 (2023) *Computer Law & Security Review* 105790.

⁵⁷Art 16(5) DSA.

⁵⁸E Douek, ‘The Meta Oversight Board and the Empty Promise of Legitimacy’ 37 (2023) *Harvard Journal of Law & Technology* 1; K Klonick, ‘The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression’ 129 (2019) *Yale Law Journal* 2418.

⁵⁹Oversight Board, ‘Q2 2023 Transparency Report’ <<https://www.oversightboard.com/transparency-reports/>> accessed 1 May 2025.

⁶⁰Art 16(5) DSA.

and-action mechanisms as a whole.⁶¹ As previously mentioned, our study demonstrates that those distrusting platforms file less notices than those that trust the same platforms. Thus, inasmuch as the DSA has the goal not to discourage the filing of notices by flaggers operating in good faith, an (at least partial) extension of the transparency obligations of Article 17 to the complainant may constitute a step in the right direction. Once again, when situating the DSA in the context of our experimental results, it becomes apparent how many provisions of this Regulation presuppose individual initiative and rely on implicit assumptions about behaviour, but sometimes fail to consider whether those presuppositions and assumptions comport with the reality of psychology, and whether the Regulation creates adequate behavioural incentives to reach the desired policy goals.

D. The impact of poster intention on reporting behaviour

A final set of DSA-related inferences can be drawn from data regarding the intention of those who post LBA content. More specifically, the study demonstrates that the participants' decision to file a notice is influenced not only by the nature of the image that they see, but also by the intention of the user that posted that image: participants reported content with a negative moral connotation more often, when that content was endorsed (rather than criticised) by its poster. These results confirm the oft-repeated assumption that content moderation is essentially contextual:⁶² especially with reference to LBA content, where no hard-and-fast prohibition exists in the law, certain posts may be perceived as acceptable when the poster wishes to draw attention to a problem, and unacceptable when the poster condones or endorses the content instead. Interestingly, this context-dependent approach is adopted by social media platforms themselves: Facebook's Community Standard on Adult Nudity and Sexual Activity, for instance, generally prohibits images of uncovered female nipples, but makes an exception when the poster shares the image 'for medical or health purposes',⁶³ such as raising awareness on mastectomy or breast cancer.⁶⁴

Furthermore, this context-specific approach to content moderation has also been embraced by the Oversight Board. For instance, the Board held that Meta should make an exception to its policy of banning videos that depict sexual harassment, when (under certain well-specified conditions) the content is newsworthy and is shared for the purpose of 'raising awareness', not 'in a sensationalised context', and without involving nudity.⁶⁵ Along similar lines, the Board held that Meta should make an exception to its Violence and Incitement Community Standard when a post contains the slogan 'death to Khamenei' (with reference to Iran's Supreme Leader, Ayatollah Khamenei), but the sentence is 'used rhetorically to mean "down with"', rather than as an actual threat of violence.⁶⁶

In sum, both the experimental results and the content moderation policies of companies like Meta (as interpreted by the Oversight Board) suggest that, when deciding whether to take a content moderation decision with regard to LBA content, the intention of the poster should not be ignored. This has two important implications for the DSA, concerning the (partial) prohibition of

⁶¹Art 24(5) DSA requires the publication of the statements of reason in the DSA Transparency Database (<<https://transparency.dsa.ec.europa.eu/>> accessed 29 March 2024). Given the breadth of this transparency obligation, it is hard to justify the exclusive focus on the 'affected recipients of the service' in Art 17.

⁶²T Gillespie, 'Content Moderation, AI, and the Question of Scale' (2020) 7 Big Data & Society 2053951720943234; T Dias Oliva, 'Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression' 20 (2020) Human Rights Law Review 607.

⁶³Meta, 'Adult Nudity and Sexual Activity' <<https://transparency.fb.com/en-gb/policies/community-standards/adult-nudity-sexual-activity/>> accessed 29 March 2024. <<https://www.facebook.com/business/help/725672454452774?id=208060977200861>>.

⁶⁴Oversight Board, *Breast Cancer Symptoms and Nudity*, 2020-004-IG-UA <<https://www.oversightboard.com/decision/IG-7THR3SII>> accessed 29 March 2024.

⁶⁵Oversight Board, *India Sexual Harassment Video*, 2022-012-IG-MR <<https://www.oversightboard.com/decision/IG-KFLY3526/>> accessed 3 June 2024.

⁶⁶Oversight Board, *Iran Protest Slogan*, 2022-013-FB-UA <<https://www.oversightboard.com/decision/FB-ZT6AJ54X/>> accessed 3 June 2024.

automated decision-making in the process of content moderation, and the regard of fundamental rights in the context of the platforms' terms of service.

From the first point of view, it is widely acknowledged that, in the practice of content moderation, many decisions are made (at least at an initial stage) algorithmically.⁶⁷ While machines may be well equipped to detect certain types of content (eg, pictures depicting nudity), they may fail to understand the intention of the poster sharing that content. For instance, an Oversight Board case revealed that, in the past, Meta's algorithms have failed to detect that a picture of uncovered female breasts had been shared to raise breast cancer awareness, and that picture had been automatically removed from the platform despite Meta's aforementioned health-related exception in its Community Standard on Adult Nudity and Sexual Activity.⁶⁸ Here, the DSA does offer some guarantees of human review. While moderation at the initial stage of notice-and-action mechanisms may be entirely algorithmic, the following stage of internal complaint handling cannot be completely automated. More specifically, pursuant to Article 20 of the DSA, following a platform's decision whether to adopt a content moderation measure or not, a complaint can be lodged by either the user posting the content, or those that submitted an (unsuccessful) Article 16 notice. When observed in the light of our experimental results, Article 20 does offer a limited but important guarantee by ensuring that a human moderator will review the poster's intention and the context within which a certain type of content was shared. A flagger that filed an unsuccessful notice will then be able to request this human review, and voice the intent-related concerns that may have induced him/her to file an Article 16 notice in the first place.

From the second point of view, the question arises whether the DSA imposes on platforms an obligation to take into account the poster's intention, when shaping their moderation policies with regard to LBA content. To be sure, the examples mentioned above demonstrate that platforms do already often consider poster intention, when determining whether a post is incompatible with their terms and conditions. The question, however, is whether the DSA obliges platforms to do so. The results of our behavioural study suggest that LBA content is perceived differently depending on the intention of the user that posted it. In light of this, it would be desirable for social media platforms to develop their terms of service in a way that accounts for poster intention, when determining to what extent lawful content should be taken down or otherwise moderated. However, the Regulation provides more questions than answers in this respect. On the one hand, article 14(4) of the DSA does refer to the fundamental rights of the poster, including freedom of expression. This reference could be construed as preventing the platforms from disregarding whether a user that posted LBA content did so with the purpose of raising awareness, participating to the public discourse, or otherwise making legitimate use of his/her freedom of expression. On the other hand, however, this provision does not apply to the drafting of terms of service that prohibit legal content, but only to the application and enforcement of those contractually imposed restrictions. Nevertheless, other DSA provisions may require platforms to structure their terms of service in a way that adequately considers poster intention. Under article 34(1)(b) of the DSA, very large platforms must subject their terms of service⁶⁹ and content moderation systems⁷⁰ to a yearly risk assessment concerning, among other factors, 'any actual or foreseeable negative effects for the exercise of fundamental rights', including the right to freedom of expression and information. In this context, it could be argued that a platform's terms of service may have a negative effect on a poster's exercise of his/her fundamental rights, inasmuch as the contractual terms treat posters with different intentions in the same way, when they post the same type of content.

⁶⁷Gillespie (n 62).

⁶⁸Oversight Board, *Breast Cancer Symptoms and Nudity*, 2020-004-IG-UA <<https://www.oversightboard.com/decision/IG-7THR3SII>> accessed 3 June 2024.

⁶⁹Art 34(2)(c) DSA.

⁷⁰Art 34(2)(b) DSA.

Finally, the insight that people's perception of content with a morally negative connotation is influenced by the intention of the user posting that content could be helpful when applying Articles 35 and 37 of the DSA. Under the former, very large platforms may be required to alter their terms of service and content moderation practices as a 'mitigation of risk' measure. This provision obliges platforms to take adequate measures, to counter the risks identified through the aforementioned Article 34 risk assessment. Relatedly, according to Article 37, very large platforms also have an obligation to subject themselves to an independent audit; in this context, the adequacy of a platform's mitigation measure can be tested. Our results suggest that the question whether a platform has failed to adequately take into account the intention of the poster, when creating and enforcing its content moderation policy, is important not only for those posting LBA content, but for a wider range of stakeholders.

4. Conclusions

The DSA is a critical piece of the procedural puzzle of generating a safe social media environment. In this article, we have highlighted how the DSA's ambitious provisions regarding content moderation rely on assumptions regarding psychology and inferences about behaviour, yet often do so in the absence of empirical behavioural evidence. The difference between how people are assumed to behave and evidence on how they actually behave is, we argue, a crucial element in the complex landscape of effectively moderating content on social media. For this reason, we shed light on the value of behavioural research in understanding the preferences, incentives and decisions of individuals, and the role of such information in interpreting and developing legal provisions. Specifically, through explaining the results of a novel experimental study designed with the DSA in mind, we offer insights on judgement and decision-making processes, and how this evidence can steer away from unfounded suppositions to help interpret and apply the DSA against the reality of behaviour, as well as inspire a future research agenda on the topic.

We discuss four aspects. Specifically, while the DSA only requires a notice-and-action mechanism with respect to illegal content, we show that the moral valence of content influences the decision whether to flag content, even if that content is not illegal. Moreover, we provide evidence supporting the importance of perceptions regarding responsibility and what the participants believe to be their role in content moderation on their ultimate behaviour to file notices. In addition, we highlight the fundamental role of perceptions and trust in platforms and how this is affected by the DSA's transparency obligations. Finally, we emphasise how participants contextualise social media content, specifically by taking on board the intention of content posters, in deciding what they think should be subject to moderation. We elaborate on what this means for social media platforms and their policies.

Ultimately, the DSA aims to create the architecture for social media platforms and content governance. However, it should not be forgotten that architecture is always meant to house people and shape individual experiences. Individuals can thus either be harmed, helped, or led astray, depending on the procedures that are available and structures that are in place for them. If the DSA truly has the goal of benefiting individuals online, it is their psychology and behaviour which ought to be explored, understood and considered.

Competing interests. This article builds on the results of an empirical study which has been published by the authors on (2024) PLOS ONE, e0300960, as acknowledged in footnote 23.