Social Media Governance Policy Brief:

How the DSA
can enable a public science of
digital platform social impacts

May 2024

GPAI / THE GLOBAL PARTNERSHIP
ON ARTIFICIAL INTELLIGENCE

Table of Contents

Summary

- A key objective of the Digital Services Act (DSA) is to provide transparency about the impacts of Very Large Online Platforms (VLOPs) on users and society more generally, and about how any harmful effects that are identified can be countered.

- A key mechanism by which the DSA proposes to achieve this is by granting independent researchers access to platforms to identify risks and harms, under a Delegated Regulation whose terms are currently being discussed.

- In order to answer critically important research questions and to identify risks, this Delegated Regulation must enable vetted external researchers to undertake ethical controlled experimentation on platforms, with appropriate consent and oversight. The gold standard of such experimentation is A/B testing.

- There is ample evidence that the results of A/B experiments on platforms can be safely disseminated, without disclosing company IP or personal user data.

- The Delegated Regulation could establish robust frameworks to uphold the rights and wellbeing of users participating in A/B tests, similar to those that govern randomised controlled trials in the medical field.

- Enabling A/B testing on VLOPs by independent researchers would foster the development of a public science of VLOP impacts on users and society.

Background: The DSA and Delegated Regulations

A key aim of the EU's Digital Services Act (DSA, 2022) is to improve transparency about the operation of very large online platforms (VLOPs): to shed light on how the algorithms and processes deployed by these platforms influence the way information flows in our society, and influence individual platform users, in potentially harmful ways. The DSA provides two particular mechanisms for delivering this transparency. One involves access to company data and processes by *external auditors*: each VLOP must undergo regular independent audits, to check for compliance on its obligations under the DSA. Another involves access to company data and processes by vetted *independent researchers,* to ensure potential risks to fundamental rights can be identified. This allows DSA-relevant aspects of company operation to be further studied, using data and methods that are only available within companies. Each type of access is governed by a Delegated Regulation. The Delegated Regulation on Auditing has already been released (DSA, 2023). The Delegated Regulation for Data Access for External Researchers is currently under discussion. Our briefing note contributes to this latter discussion.

In this note, we argue that the Delegated Regulation for Data Access to Researchers must empower researchers to run *controlled experiments* within VLOPs—in particular **A/B tests**, which companies themselves often use to guide system development. We argue that researchers cannot gain an adequate understanding of platform operation through passive observation: they must be able to *actively explore* options available to VLOPs, to fully understand the systems VLOPs deploy, and the effects these systems can have on users, and on society as a whole. This is what A/B tests permit. Scientific research is an active process, that advances through experiments. If the DSA is to support a *public science* of VLOP effects, external researchers need far more than the summary data companies currently provide: they need access to the experimental paradigms that VLOPs make use of themselves. A/B tests are the gold standard experimental method for measuring platform effects on users (e.g., Walker et al, 2019): they should be a cornerstone for DSA-enabled external research.

Enabling systemic risk assessment for VLOPs

A priority for the DSA is to assess possible *risks* posed to society by the largest online platforms—and to mitigate any risks that are found (DSA 2022, Recitals 79-90). To meet these objectives, external researchers granted access to VLOPs under the DSA must have the tools to study these risks, and to explore how any risks that are found can be mitigated. Central among these tools is the ability to run experiments, of the kind that VLOPs run themselves. Assessing the risks of a platform service on users is impossible without controlled experiments that *vary* this service: it is only through such experiments that we can learn which risks arise *due to the platform*. Through experiments that subtly vary services offered to users, researchers will be able to effectively test many hypotheses that have been advanced about harms arising from VLOPs, such as harms to mental health in teenagers (see e.g. Twenge et al., 2022), harms to body image and eating behaviour in women (see e.g. Zeeni et al., 2023), and social harms arising from exposure to toxic content (see e.g. Baker et al., 2024). Identifying how to mitigate any risks identified also requires experiments that empirically test the effectiveness of mitigation techniques.

VLOPs have the technical ability to conduct experiments of all these kinds. But because their priorities are commercial, driven mainly by advertising revenue, they are not strongly incentivised to conduct testing on risks and harms to users, or to publicly report on such testing if they do conduct it. The DSA needs to allow external researchers to use these same experimental methods, to conduct studies focussing on the public good.

---

[1] Photo of [Adem AY](#) on [Unsplash](#)

A/B tests: The methodological gold standard for understanding platform impacts
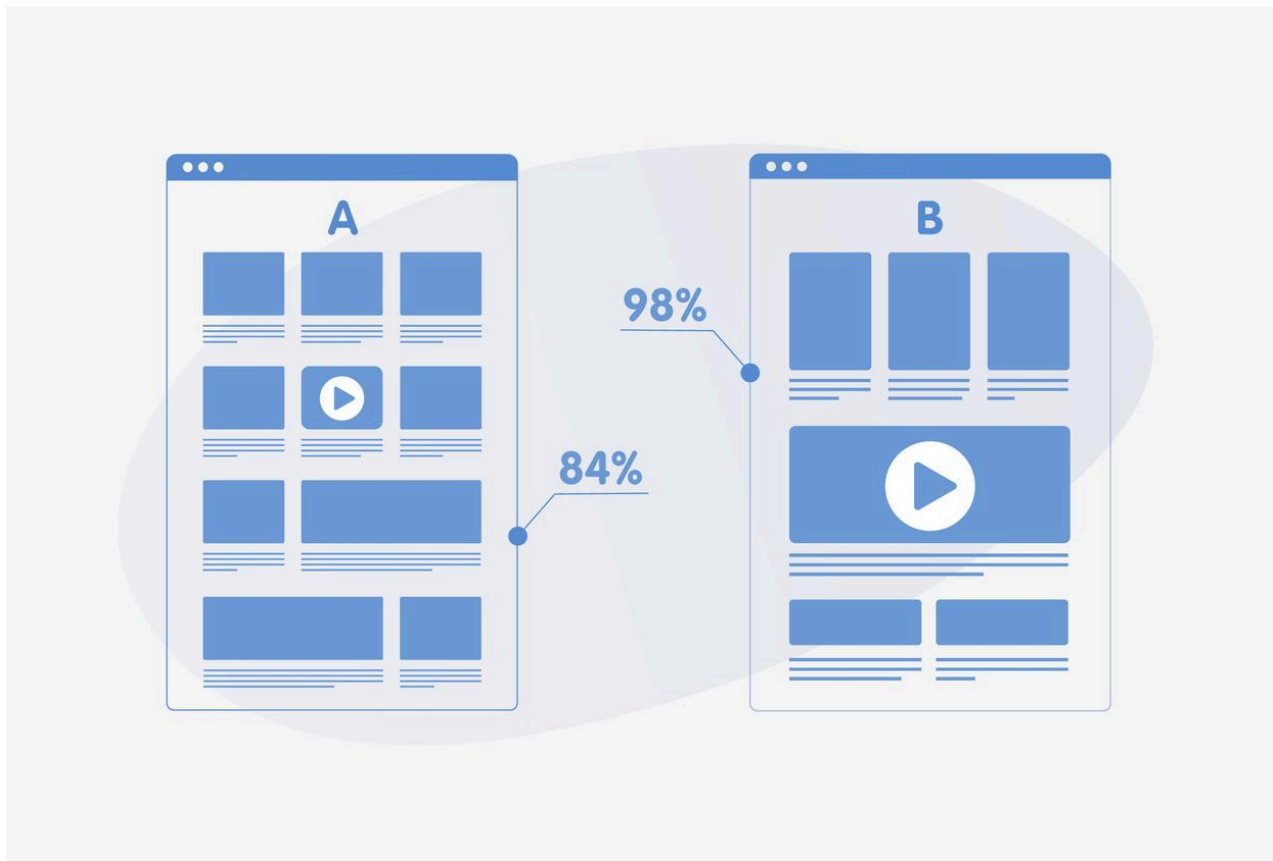
An A/B test, or Randomised Controlled Trial (Kohavi et al., 2020; Austrian et al., 2021) is a way of exploring the effects of *different versions* of a company system on its user base. The procedure for an A/B test can be simply stated. We first create a set of *user groups*, by sampling randomly from some user population. Groups may represent the population as a whole, or some targeted population, such as teenagers. Each group is then given a different version of the system: the different versions should vary in a single element or feature of the system. The behaviour of users in each group is then monitored and recorded over some trial period, as they interact with their allocated version of the system. If, after this trial period, there are any significant differences between groups, these can be reliably attributed to the different systems they interacted with—because the only systematic difference between the groups, statistically speaking, is the system they engaged with. In fact, we can confidently say the observed differences in user behaviours across groups were *caused* by the differences in the systems they were given.

A/B tests provide valuable concrete information about what user effects are *within the control* of companies. They intervene in specific aspects of the user experience, and

precisely measure the effects of these interventions. They are controlled experiments, in that they systematically vary one aspect of the user experience. They can operate at large scales—for instance, Twitter ran an A/B test on its recommender algorithm that monitored 5% of its global user base, over a period of several years (Huszár et al., 2021).

As already noted, A/B tests are the gold standard method for empirically assessing the effects of VLOP systems on users. They have surfaced many important findings: for instance, a 2021 Instagram A/B test showed that hiding 'like counts' of user posts reduced user 'negative social comparison' scores by 2% (WSJ, 2021); Huszár et al.'s 2021 Twitter study showed that a recommender system that learns from user behaviour presents users with vastly more political content than one that just presents feed items in reverse-chronological order. Our group of international experts at GPAI has argued in several reports that company A/B tests provide by far the best empirical methodology for studying the effects of social media recommender systems (see e.g. GPAI, 2021; 2022). They overcome problems with sampling bias and confounding variables that routinely affect experiments run externally to companies. Crucially, because they intervene in the user experience, they also provide the only way of effectively testing *causal hypotheses* about the effects of recommender algorithms.

A proposal: DSA-governed researcher access to A/B tests

[2]

Our main proposal here is that the DSA Delegated Regulation on Data Access should enable external researchers to conduct experiments on the platforms they access. In particular, they should allow external researchers to conduct A/B tests, using the methods standardly available to companies.

Safety

We have already argued that external researchers can *safely* use company A/B testing methods, without compromising company IP or user privacy (see again GPAI 2021; 2022). In relation to IP, A/B tests are 'black-box' techniques: they can be conducted without access to the code of the algorithms being tested. (The transparency they provide is about the *effects* of algorithms, rather than their internal operation. Crucially, effects are the central concern for the DSA.) In relation to user privacy, A/B tests summarise *high-level measures* of user behaviour, aggregated over *large user groups:* so there is no disclosure of private information of individual platform users. Useful validation for these arguments comes from work conducted under the Christchurch Call Initiative on Algorithmic Outcomes. This initiative explicitly focussed on the construction of a privacy-enhancing interface for researchers to company A/B testing infrastructure (see ChCh 2022).

Best practice and ethics review

VLOPs already have a code of best practice for A/B tests, as documented in a recent report by the Integrity Institute, squarely focussed on the DSA (Allen and Lawson, 2024). To test a given hypothesis about user effects, best practice is to begin with 'offline' studies that test the hypothesis in a simulated version of the platform, which runs without real users. If these studies reveal effects requiring further attention, a preliminary A/B test can be trialled on company-internal evaluators, to preview the likely user experience. If the user experience is deemed acceptable, a full A/B test can be conducted. We envisage the Delegated Regulation on Data Access would require a progressive series of safety checks of this kind. (Note this means the Regulation should enable external researchers to run tests on company-internal simulation platforms, where these are available.) The Regulation must also put a more rigorous ethical review process in place, to oversee any research that is conducted. There is a growing literature on the ethics of A/B tests (see Polonioli et al., 2023 for a survey), which regulators can draw on here, to extend companies' existing practices, which often do not include an explicit ethics review.

User consent

When an A/B test is run on real users, a critical requirement is to obtain their *consent*. For A/B tests run by companies, for their own commercial purposes, consent is often built into the terms and conditions users agree to. This default consent has proven sufficient for some of the A/B studies that have been reported publicly—for instance, that of Huszár et al. (2021). But more recent A/B studies have run on subjects who explicitly consented (see

e.g. Guess et al., 2023). We suggest explicit user consent would be vital for researcher-initiated A/B tests conducted under the framework of the DSA.[3]

A governance framework for safe and ethical A/B tests

Experiments in any domain that overtly explore 'harmful effects' must be overseen with great caution. But the idea of tests that measure harm is hardly new territory for regulators. The domain of medical testing is an obvious reference point to consider. Many countries have processes for testing novel medical treatments, in randomised controlled trials that are methodologically very similar to the A/B tests we propose. Tests are interventions, that trial products developed by commercial companies, on consenting subjects: these tests are publicly pre-registered, and their results are made public. They are run by researchers without competing commercial interests, or who declare possible conflicts of interest. Tests are closely overseen, to ensure there is no compromise to the privacy, rights and wellbeing of participating subjects, or the IP of companies. They provide a carefully designed mechanism for quantitatively assessing the costs and benefits of treatments, which ultimately serves the public good. Each proposed trial is rigorously assessed, to ensure that only well motivated and ethically justifiable trials are conducted. The A/B testing processes we are proposing for external researchers under the DSA fulfil a similar function, in the realm of technology effects on citizens.

External researchers' ability to conduct A/B tests on company platforms under the DSA would have to be subject to tight controls. The number of A/B tests performed will need to be limited, because tests consume resources, and intervene in user experiences. A rigorous process would be needed for soliciting *proposals* for A/B tests to be conducted under the DSA, and rigorously *evaluating* these proposals, to choose those which are best motivated scientifically, and which promise to deliver the most informative and socially beneficial results. Proposals would also be assessed on ethical grounds, like all scientific research proposals. (As already noted, there are opportunities to improve companies' current practice.) Finally, the commercial independence of those undertaking the research should be clearly established. We envisage the review process would operate much like the calls for research proposals presently conducted by public research funding bodies.

Conclusion

---

[3] These two studies also illustrate a relevant methodological difference in A/B testing. Huszár et al.'s study *reanalysed* the results of an A/B test that had already been conducted by Twitter for its own commercial purposes, while Guess et al.'s study was a *new* A/B test conducted to test a specific hypothesis (as far as we understand). Consent for Huszár et al.'s study was covered by Twitter's standard terms and conditions, which gave the company license to conduct experiments 'for the business purpose of improving the [recommendation] algorithm'. Guess's study was conducted for a purpose not covered by Facebook's terms and conditions; this may be why explicit consent was sought, but we may also just be moving towards more explicit consent procedures. In any case, external researchers operating under the DSA should be able to use both of these methods: reanalysis of logs from existing company A/B tests, or if appropriate, running of new A/B tests.

With appropriate governance and consent procedures in place, we believe external researchers can safely conduct A/B tests on platform users—and we recommend that the Delegated Regulation on Data Access should provide external researchers with the ability to conduct A/B tests. In fact, we believe this provision for external researchers is essential support for a provision already granted to auditors, in the Delegated Regulation for Audits, to conduct 'experiments'.[4] Auditors are not researchers: any experiments auditors conduct on a platform will have to follow methodologies established by prior research, so the DSA as a whole needs to empower researchers to identify appropriate experimental methodologies, through experiments of their own.

We believe that by the mechanisms discussed in this brief, the DSA can function to institute a new *public science of VLOP impacts*, operating safely within the large private companies that provide infrastructural tech services to EU citizens. The scope of this new science, while technically limited to the EU jurisdiction, will spread beyond the EU, in disseminating information about platform effects to other jurisdictions.

To reiterate: the research enabled by the DSA cannot take as its starting point the data that VLOPs currently make available. It must start from the compelling questions of public interest that lie at the foundation of the Act: do VLOPs pose risks to individual users, and to society? If so, how can these be mitigated? The only way to start from these key questions is to enable external researchers to conduct A/B tests.

References

Allen, J and Lawson, A. (2024). On risk assessment and mitigation for algorithmic systems. Integrity Institute report.
https://drive.google.com/file/d/1ZMt7igUcKUq00yakCnbxBCcaA7vajAix/view

J. Austrian, F. Mendoza, A. Szerencsy, L. Fenelon, L.I Horwitz, S. Jones, M. Kuznetsova, D.M. Mann, Applying A/B Testing to Clinical Decision Support: Rapid Randomized Controlled Trials, J Med Internet Res 2021;23(4):e16651, doi: 10.2196/16651

Baker, C., Ging, D. and Brant Andreasen, M. (2024). Recommending Toxicity: The role of algorithmic recommender functions on YouTube Shorts and TikTok in promoting male supremacist influencers. DCU Anti-Bullying Centre, Dublin City University.
https://antibullyingcentre.ie/wp-content/uploads/2024/04/DCU-Toxicity-Full-Report.pdf

ChCh (2022). Christchurch Call Initiative on Algorithmic Outcomes. New Zealand government press release.
https://www.beehive.govt.nz/release/christchurch-call-initiative-algorithmic-outcomes.

---

[4] Article 2(17) of the Delegated Regulation on Auditing (DSA, 2023) defines the 'tests' auditors can conduct. The definition includes 'measurements, *experiments* or other checks, including checks of algorithmic systems' (our emphasis).

DSA (2022). 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).

DSA (2023). Delegated Regulation on independent audits under the Digital Services Act. https://digital-strategy.ec.europa.eu/en/library/delegated-regulation-independent-audits-under-digital-services-act

GPAI (2021). Responsible AI for Social Media Governance A proposed collaborative method for studying the effects of social media recommender systems on users. Global Partnership on AI report.

GPAI (2022). Transparency Mechanisms for Social Media Recommender Algorithms: From Proposals to Action Tracking GPAI's Proposed Fact Finding Study in This Year's Regulatory Discussions. Global Partnership on AI report.

Guess, A. et al., How do social media feed algorithms affect attitudes and behavior in an election campaign? Science 381, 398-404 (2023). https://www.science.org/doi/10.1126/science.abp9364

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on Twitter. Proceedings of the National Academy of Sciences, 119(1), e2025334119.

Kohavi R, Tang D, Xu Y. Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. Cambridge: Cambridge University Press; 2020.

Polonioli, A., Ghioni, R., Greco, C., Juneja, P., Tagliabue, J., Watson, D., & Floridi, L. (2023). The Ethics of Online Controlled Experiments (A/B Testing). Minds and Machines, 1-27.

Twenge, J. M., Haidt, J., Lozano, J., & Cummins, K. M. (2022). Specification curve analysis shows that social media use is linked to poor mental health, especially among girls. Acta Psychologica, 224, 103512.

Walker et al   "Top Challenges from the first Practical Online Controlled Experiments Summit". SIGKDD Explorations. 21 (1): 20–35, June 2019. doi:10.1145/3331651.3331655

WSJ (2021). Social comparison: Topics, like counts, celebrities, selfies. Facebook research document, leaked to the Wall St Journal. https://s.wsj.net/public/resources/documents/social-comparison-topics-celebrities-like-counts-selfies.pdf

Zeeni, N., Abi Kharma, J., & Mattar, L. (2023). Social media use impacts body image and eating behavior in pregnant women. Current Psychology, 42(6), 4948-4955.