Expert Submission for UNESCO's Guidance for Regulating Digital Platforms: a Multistakeholder Approach

Submission

Final Version – 20 January 2023

Prepared by:

By Dr Mark R. Leiser (VU-Amsterdam)[1]
Ms Maria Rebrean (Leiden University)[2]
Dr Philipp Lorenz-Spreen (Max Planck Institute for Human Development, Berlin)[3]
Professor Ralph Hertwig (Max Planck Institute for Human Development, Berlin)[4]
Dr Stefan M. Herzog (Max Planck Institute for Human Development, Berlin)[5]
Dr Anastasia Kozyreva (Max Planck Institute for Human Development, Berlin)[6]
Dr Sabine K. Witting (Leiden University)[7]

[1] Dr Mark Leiser is a Professor in Digital, Internet, and Platform Regulation at Vrije University of Amsterdam, with expertise in fundamental and human rights.
[2] Maria Rebrean is an Advanced Masters Student in Law and Digital Technologies at Leiden University
[3] Dr Philipp Lorenz-Spreen is Research Scientist in Adaptive Rationality at the Max Planck Institute for Human Development, Berlin.
[4] Professor Ralph Hertwig is the Director of the Center for Adaptive Rationality at the Max Planck Institute for Human Development, Berlin.
[5] Dr Stefan Herzog is a Senior Research Scientist and Head of Research "Area Boosting Decision Making" | Center for Adaptive Rationality (ARC), Max Planck Institute for Human Development, Berlin.
[6] Dr Anastasia Kozyreva is Research Scientist in Adaptive Rationality at the Max Planck Institute for Human Development, Berlin.
[7] Dr Sabine Witting is Assistant Professor at eLaw - Center for Law and Digital Technologies

UNESCO's Global Guidance for Regulating Digital Platforms for Information as a Public Good

We are a group of academics, lawyers, and scientists concerned with the impact digital technologies have on both individuals and societies. We have been collectively working via a research grant provided by the Volkswagen Foundation on the impact of algorithms, AI, and the business models of platforms on users' abilities to make good decisions about the information and environments they routinely encounter online.[8] The project's objective is to identify evidence-based ways to reclaim individual autonomy and to redress the imbalance in the relationship between human decision-makers and platforms; for example, by investigating what impact social media has on democratic values, what are people's attitudes to key aspects of digital architectures and content moderation policies, and how to combat online misinformation and manipulation while respecting fundamental and human rights. UNESCO's Guidelines for Regulating Digital Platforms provide an opportunity to develop a regulatory framework for effective decision-making regarding the quality of information encountered in a manner that respects fundamental and human rights while empowering users. In this submission, we argue that any regulatory regime addressing the role of platforms in combating the spread of illegal or harmful content must take an evidence-based approach to the transparency required to ensure platforms enhance the availability of accurate and reliable information to the user.

With systemic risks and harms justifying increased interest in regulating platforms, several jurisdictions have taken measures to increase their accountability. Most notably, the European Union's Digital Services Act[9] attempts to protect users from the harmful effects of social media by imposing a mix of reporting and transparency obligations that respect users' fundamental rights and freedoms[10] To achieve this end, it views the entire digital ecosystem holistically and supports platform growth while encouraging innovation in the delivery of services.[11] The DSA represents a technologically neutral framework, specifies rules on due diligence obligations and implements enforcement mechanisms[12], alongside safeguards that combat the harmful impacts stemming from self-regulating platforms while protecting the fundamental rights of citizens.[13] Much of the prevailing narrative surrounding rethinking platform regulation emphasises managing the risk of abuse of the platform's functionality and the societal harms commonly associated with amplifying disinformation and other harmful content, with a limited understanding of how proposed regulatory regimes actually impact human behaviour. However, behavioural science is having an impact on platform regulation in Europe. The DSA is informed, in part, by behavioural science, but these efforts could be stepped up further.[14]

Unfortunately, many digital platforms and associated services rely heavily on transparency mechanisms[15] without sufficient evidence to demonstrate that these mechanisms effectively inform and empower users to make informed decisions about the quality of the information they encounter.[16]

Therefore, we recommend UNESCO embraces insights from behavioural sciences regarding the design of environments that empower users, the measurement of the effectiveness of transparency obligations, and user preferences for personalised recommender systems and targeted advertisements

[8] https://sks.to/rao

[9] 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) https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065&from=EN

[10] Recital 3, Digital Services Act.

[11] Recitals 1&4, Digital Services Act.

[12] Article 1, Digital Services Act.

[13] Article 13, Digital Services Act.

[14]Lewandowsky, S., Smillie, L., Garcia, D., Hertwig, R., Weatherall, J., Egidy, S., Robertson, R.E., O'Connor, C., Kozyreva, A., Lorenz-Spreen, P., Blaschke, Y. and Leiser, M., Technology and Democracy: Understanding the influence of online technologies on political behaviour and decision-making, EUR 30422 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-24089-1, doi:10.2760/593478, JRC122023.

[15] Article 29, Digital Services Act.

[16] Leiser, M. (2016). The problem with 'dots': questioning the role of rationality in the online environment. *International Review of Law, Computers & Technology*, *30*(3), 191-210.

over one-size-fits-all cookie notices.[17] Accordingly, our submission reflects on four of the five principles set out in UNESCO's guidance for regulating platforms and how insights from the behavioural and social sciences can help inform effective and meaningful platform regulation.

21.1 Platforms have content governance policies and practices consistent with human rights standards, implemented algorithmically or through human means (with adequate protection for the well-being of human moderators);

Human Rights

In 1991, UNESCO produced the Windhoek Declaration, furthering the right to free expression.[18] The influential declaration emphasizes information as a tool for the furthering of fundamental rights, including "democratic governance and sustainable development, leaving no one behind".[19] We usually associate the dissemination of information as a fundamental right or as instrumental to a fully functional 'marketplace of ideas'. However, this oft-cited phrase attributed to Justice Oliver Wendell Holmes is abridged and misquoted. The actual quote is, "the ultimate good desired is better reached by free trade in ideas–that the best test of truth is the power of the thought to get itself accepted in the competition of the market, and that truth is the only ground upon which their wishes can be carried out". In other words, the marketplace of ideas is worth protecting only when ideas contribute to finding the truth.

The information ecosystem is in crisis. The amplification and ease of dissemination have activated fringe elements of society, making dangerous narratives seem popular. While measures ensuring the fundamental right to expression should always be in the foreground of effective platform regulation, disinformation spread via computational propaganda could also interfere with the right to free elections under Article 3 of Protocol No. 1 of the Convention. Getting the balance right will be the battleground of the 21st Century.

Moreover, unsolvable debates rage about the role of platforms and their responsibility to protect fundamental rights; for example, is it within the public's interest to receive false information that furthers freedom of expression at the expense of the very rights that the dissemination of information is supposed to further? Platforms are not only of fundamental importance to free expression but are instrumental in the proper regulation of false information. They have an instrumental role in the amplification and dissemination of falsehoods. Leaving the responsibility of policing content to self-regulating private parties shielded in a "safe harbour" until they gain knowledge of problematic content is no longer seen as acceptable.

Discussions about platforms' role as gatekeepers are often normative and fail to address how content is actively managed, users are profiled, and how the delivery techniques are unique to each platform. Nor do they properly acknowledge insights from the behavioural sciences that suggest how information is presented and received and can have significantly different outcomes on how that information is understood. For example, Facebook's decision to black out content flagged by trusted fact-checkers[20] significantly reduced content virality. Although users access it, they do not share the material for fear of social embarrassment. In this sense, the freedom of expression of the content creator was preserved, but the effect of amplification was modified by a better understanding of human behaviour.

Targeted advertising also exists at the intersection of commerce, consumer and data protection, online manipulation and human rights abuses. Targeted advertising's invasive and opaque use of personal data, for example, raises issues for the right to privacy, as the 'protection of personal data is of fundamental importance to a person's enjoyment of his or her right to respect for private and family

[17] See insights provided by, for example: Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021). Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the US. Humanities & Social Sciences Communications, 8(117). https://doi.org/10.1057/s41599-021-00787-w

[18] UNESCO, Windhoek + 30 Declaration: information as a public good, World Press Freedom Day 2021, (2021).

[19] Id at para 7

[20] See for example: Pasternack A., How Facebook pressures its fact-checkers., Fast Company. (2020). Last accessed: 20 January 2023. https://www.fastcompany.com/90538655/facebook-is-quietly-pressuring-its-independent-fact-checkers-to-change-their-rulings

life'.[21] Platforms' obligations about ad transparency are limited to obligations about informational disclosures ('Why am I seeing this ad?' or 'Sponsored Ad'). Obligations rarely consider insights from the behavioural sciences about how users want the information to be presented to them, nor evidence of their actual interaction.

Summary of Main Discussion Points

- UNESCO produced the Windhoek Declaration in 1991, which furthers the right to free expression and emphasizes information as a tool for furthering fundamental rights such as democratic governance and sustainable development
- The phrase "marketplace of ideas" is often cited as a reason for protecting free expression, but it is important to remember that this concept only holds value when ideas contribute to finding the truth
- The current information ecosystem is in crisis, with the amplification and ease of dissemination of information allowing dangerous narratives to spread and potentially interfering with the right to free elections
- Discussions about platforms' role as gatekeepers of information are often normative and fail to address the active management of content, user profiling, and delivery techniques unique to each platform
- Platforms also have an instrumental role in the amplification and dissemination of false information and the responsibility for policing content should no longer be left to self-regulating private parties
- Targeted advertising raises issues for the right to privacy and platforms' obligations about ad transparency are often limited to informational disclosures and do not consider insights from behavioural sciences about how users want information to be presented to them.

[21] Article 8, ECHR.

21.2 Platforms are transparent, being open about how they operate (taking into account commercial confidentiality) with policies being explainable

Transparency as a Legal Concept

Regulatory regimes across the globe have faced many challenges regarding technology regulations. A significant contributing factor is the convoluted nature of technological development and processes. As a result, regulators have established regulatory regimes that centre around the concept of transparency. For example, the DSA[22] and the GDPR[23] contain several articles designed to establish transparency between platforms and consumers. Despite the heavy reliance of digital regulatory measures on the principle of transparency, the term is only vaguely defined.[24]

In practice, and as will be further elaborated, transparency requirements resulted in often incomprehensible, large bodies of technical or legalist text.[25] The resulting transparency fallacy (i.e., transparency in name only) became an infamous characteristic of European digital legal regimes in Europe. Criticising the legal approach to transparency, Weatherill[26] writes: “Their [transparency techniques] efficacy depends on the capacity of the consumer to process the information that is supplied and to act rationally in response to it. In so far as consumers fail to behave in an alert and rational manner, regulatory intervention based on information disclosure may not yield the intended benefits. In reality, some consumers will typically not be alert; others will not be capable of being alert.”

Transparency Failures

Increased legislative efforts to require platforms to provide ‘transparency’ have brought an onslaught of opaque privacy policies[27] and endless displays of consent-based ‘cookie’ notices. These displays of legalistic text, emphasized by all-or-nothing user choices, have elicited uncalculated human reactions and ‘consent fatigue’. The situation has gotten so bad that users display either *rational ignorance*[28] (“When the cost of acquiring information is greater than the benefits to be derived from the information, it is rational to be ignorant”[29]) or rational *inattention*. In other words, due to time constraints, it is in the user’s interest not to read the privacy notices, terms and conditions, etc. Instead of closing knowledge gaps, making informed decisions, and increasing user autonomy, the concept of transparency, in its present state, is responsible for platforms’ pursuit of commercial objectives while ignoring wider responsibilities to users and society alike. Legal requirements for transparency might bring about comprehensive privacy policies and consent pop-ups, but they do not empower users to make informed decisions about the manner of their digital engagement.[30]

[22] Articles 24 and 29, DSA

[23] Articles 5 and 12-14, GDPR.

[24] See also: Andrada G., Clowes, R.W., & Smart, P.R. "Varieties of transparency: exploring agency within AI systems." AI & society (2022): 1-11.

[25] Brunotte, W., Chazette, L., Kohler, L., Klunder J., & Schneider K.,. "What About My Privacy? Helping Users Understand Online Privacy Policies." Proceedings of the International Conference on Software and System Processes and International Conference on Global Software Engineering. (2022).

[26] Weatherill,S., EU Consumer Law and Policy, Elgar European Law series (2013), p. 93.

[27] Brunotte, W., Chazette, L., Kohler, L., Klunder J., & Schneider K.,. "What About My Privacy? Helping Users Understand Online Privacy Policies." Proceedings of the International Conference on Software and System Processes and International Conference on Global Software Engineering. (2022).

[28] https://www.theguardian.com/technology/2017/mar/03/terms-of-service-online-contracts-fine-print

[29] Downs, Anthony. "An economic theory of democracy." (1957): 260-276.

[30] Lorenz-Spreen, P., Geers, M., Pachur, T., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021). Boosting people’s ability to detect microtargeted advertising. Scientific Reports, 11(1), 1-9; Andreou, A., Venkatadri, G., Goga, O., Gummadi, K. P., Loiseau, P., & Mislove, A. (2018, February). Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations. In NDSS 2018-Network and Distributed System Security Symposium (pp. 1-15).

Design Transparency

'Dark patterns' is a term commonly used by the web collective to describe a deceptive user interface that exploits users into doing something they would not normally do.[31] It is a coercive and manipulative design technique used by web designers when some sort of action is needed from a user - typically to begin the processing of personal data or indication of agreement to a contract.[32] The user interface is the part of a system that the user interacts with and perceives. Besides what the user perceives, there is code that provides the functionality of which the user can only guess how it works. Some refer to it as the "inner mechanics", but others refer to it as the "software", the "system behaviour", the "algorithm", the "back end", the "code", or the "system architecture". Users develop mental models to describe how something works, but they do not really know what is going on inside a system. Even the most informed system engineer's mental model will contain gaps and abstractions in their understanding. A user's mental model is limited to prior experience, education, trust, etc. Users only perceive the software's functionality through the interface to which they are exposed. The current regulation of dark patterns focuses on the user interface rather than the entirety of the system.

Dark patterns compromise legal requirements like consent and privacy-by-design and legal principles like fairness and transparency. While Article 25 DSA suggests there is a prohibition on their use, the language adopted in the DSA is already insufficient for their proper regulation. The combination of deceptive design and AI-powered personalization of decision environments suggests we are in a new era of manipulative techniques used to benefit platforms at the expense of user autonomy. While some of the DSA's measures place control in the user's hands, other interventions are purely protectionist in nature; for example, Recital 67 DSA requires online marketplaces "not seek to subvert or impair the autonomy, decision-making, or choice of the recipient of the service through the design, structure, function or manner of operating of their online interface". The DSA fails to cover the inner mechanics of the system and whether those inner mechanics are behaving in an unfair/manipulative/deceptive way; for example, "Providers of online platforms shall not design, organise or operate their *online interfaces* in a way..." [Emphasis Added]. Similarly, the phrase "functionalities of an online interface" found in Recital 67 would require a generous interpretation to include the 'inner mechanics'. "Functionality" would require a court to understand how a system behaves and examine the range of actions or operations that can be performed.

Dark patterns, like most technological forms of manipulation, are evolving faster than the law can keep up. Self-evident dark patterns, like confirm-shaming, are visible to the consumer. Nagging, for example, is in your face and fairly unmissable. Other dark patterns are hidden in the user interface. Sneaking and dark misdirection patterns are designed to go unnoticed by the user. Multi-page business logic dark patterns are based on a simple algorithm (e.g. imagine a multi-step questionnaire where you answer a few questions and some branching logic drives you to one offer or another based on your responses). These types of dark patterns are simple and deterministic and mappable in a flowchart.

Complex deterministic dark patterns cannot be intuited by looking at a flow chart but will always give the same output based on the same inputs. This is typical of recommender systems. Stochastic (non-deterministic) dark patterns exist in a system, using a black box that nobody can precisely explain why the system output what it did, and the system may give different outputs to the same inputs (e.g. AI, ML, ML systems, etc.). Complex deterministic and stochastic (non-deterministic) dark patterns cannot be identified, even by an expert. They require inside information on the inner workings of the design interface and the system architecture. Platform regulation should reflect these new forms of dark patterns and require design transparency, alongside auditing by an enforcement agency.

[31] Leiser, M. R., & Caruana, M. (2021). Dark Patterns: Light to be found in Europe's Consumer Protection Regime. Journal of European Consumer and Market Law, 10(6), 237-251.

[32] Leiser, M. R. (2022). Dark patterns: The case for regulatory pluralism between the European Unions consumer and data protection regimes. In Research Handbook on EU Data Protection Law (pp. 240-269). Edward Elgar Publishing.

Algorithmic Transparency

Conversely, lawmakers have increasingly relied on the principle of transparency as the legal justification for untangling and explaining the inner workings of algorithmic black boxes and AI systems. Algorithmic transparency remains an essential tool in the regulator's arsenal, but it does not help users to take informed action against a decision made by an algorithmic system.[33] For example, Article 29(1) DSA requires platforms to describe within terms and conditions their proprietary recommender systems alongside ways users can adapt to their preferences. To adjust a platform's recommender system, users must navigate platforms' privacy policies, understand the explanatory information provided, and learn not only what areas of the system can be changed but also how. Thus, the DSA attempts user empowerment under the veil of choice without consideration of whether the user *understands* the algorithm of the decision. Annany and Crawford challenge narratives surrounding algorithmic transparency: "it assumes not only the legibility of information but also the competence of audiences to interpret and leverage information as an instrument of accountability".[34] Algorithmic transparency, in its present state, is based on a fallacy of assuming a direct link between visibility and understanding.

Meaningful Transparency

As platforms and their power evolve, so should the concept of transparency. Meaningful transparency describes the disclosure process of the most relevant, targeted information that would benefit the intended recipient.[35] In relation to technological developments, information qualifying under the label of meaningful transparency would be that which generates

(1) the user's action and understanding of[36]

(2) accountability[37]

Accountability covers both the regulator-platform and user-platform relationship. Although equally important, the first category tends to be misrepresented within regulatory frameworks[38] that either over- or under-estimate the individual's ability to interact with given information.

The term 'meaningful' is consistently employed within supranational transparency requirements[39] particularly regarding processes that include automated decision-making systems.[40] However, "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject"[41] were not translated into targeted and relevant information for the recipient. Rather, as previously indicated, meaningful transparency was interpreted as unabbreviated and incomprehensible privacy documents. A few studies from the behavioural sciences in the areas of profiling and content moderation shed light on how meaningful transparency can be applied to empower users.

For example, Kim *et al.'s* study[42] on the effects of ad transparency suggested people make better decisions when provided with explanations about targeting practices. If given details about collecting their personal data and how the platform makes inferences about them, users are better equipped to

[33] Gillis, T., and Simmons, J. "Explanation < Justification: GDPR and the Perils of Privacy." Pennsylvania Journal of Law and Innovation, 2019, pp. 71-99.

[34] Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973–989. https://doi.org/10.1177/1461444816676645

[35] Norval, C, et al. "Disclosure by Design: Designing Information Disclosures to Support Meaningful Transparency and Accountability." FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, https://dl.acm.org/doi/fullHtml/10.1145/3531146.3533133#BibPLXBIB0002.

[36] Stohl, C., et al. "Managing Opacity: Information Visibility and the Paradox of Transparency in the Digital Age." FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2016, pp. 123-137.

[37] Gillis, T., and Simmons, J. "Explanation < Justification: GDPR and the Perils of Privacy." Pennsylvania Journal of Law and Innovation, 2019, pp. 71-99. → **SEE ALSO ABOVE**

[38] Id.

[39] Article 12-14, GDPR.

[40] Article 22, GDPR; See also Wachter, S., Mittelstadt B., and Russell C. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." Harv. JL & Tech. 31 (2017): 841.

[41] Article 14, GDPR.

[42] Kim, T., Barasz, K., & John, L. K. (2019). Why am I seeing this ad? The effect of ad transparency on ad effectiveness. *Journal of Consumer Research*, *45*(5), 906-932.

make informed decisions about the targeted advertisement. This study suggests that if platforms would provide users with the right options, people will exercise more specific preferences. Current interpretations of transparency merely reflect what the platform wants us to know about their business practices, whereby meaningful transparency should reflect what the user wants to know. Thus, an environment that does not give consumers the information they desire does not amount to meaningful transparency.

Current narratives around content moderation practices place emphasis on the importance of platforms as gatekeepers - responsible for the promotion of the right to free expression and, in some instances, the right of ‘reach’ - the amplification of certain content and the deamplification of certain users (shadow banning). This presumption is included in legislative instruments like the DSA, which creates legal obligations for transparency about decisions to moderate content, extending to rights of appeal for content removal decisions (see also Meta’s Oversight Board).

Absent from this narrative is the role of individuals in decisions about the content they see. Kozyreva et al.’s study of user’s responses to various degrees of disinformation suggests users, when asked to provide judgement about how they would personally moderate content, show relatively specific preferences (e.g., for minimising harm, punishing repeated offences, and considering the reach of the offender). This study demonstrates that users choose to moderate content differently when confronted with a variety of scenarios in a controlled setting.[43] These preferences often contradict the content moderation policies of platforms, which lack transparent and user-oriented customization options. As such, platforms should integrate transparency tools enabling individuals to indicate their preferences regarding various content. Content moderation on social media is a complex problem which involves difficult moral decisions between values. So far, social media platforms have acted as the sole decision-makers on the matter, effectively putting them as arbiters of free speech. No matter what side they take, it is problematic that commercial companies or their CEOs are making decisions about speech governance.

What is needed is a consistent cross-platform approach to the governance of online speech. To design such rules, several factors and actors should be taken into account. People's preferences are not the only benchmark for making important trade-offs on content moderation. However, ignoring their preferences altogether risks undermining the public's trust in content moderation policies and regulations. Results such as those reported in Kozyreva *et al*.’s study can contribute to the process of establishing transparent and consistent rules for content moderation that are generally accepted by the public.

All of the above examples suggest a more dynamic approach to platform regulation is needed. Keeping the status quo in place will only ensure disinformation and harmful content will flow, despite evidence that users would judge content differently if provided with a choice to do so. The issue is that the advancement of user autonomy and control is subject to the business models of platforms that emphasize quantitative dissemination against the qualitative value of good content. We propose a solution based on platforms improving their social function by providing effective user-centric content moderation practices.

Finally, a clear example of the difference between nominal transparency and empowerment comes from an experiment that taught people to spot advertisements targeting them based on their personality. The results showed that a simple disclaimer did not help participants identify advertisements, but a prompt to actively reflect on their personality significantly improved their ability to spot targeted ads. This is another example of the gap between legal/descriptive transparency and meaningful transparency that empowers users to put information into action.[44]

[43] Kozyreva, Anastasia, Stefan M. Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser, and Jason Reifler. 2022. “Free Speech Vs. Harmful Misinformation: Moral Dilemmas in Online Content Moderation.” PsyArXiv. June 18. doi:10.31234/osf.io/2pc3a.

[44] Lorenz-Spreen, P., Geers, M., Pachur, T., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021). Boosting people’s ability to detect microtargeted advertising. Scientific Reports, 11(1), 1-9.

A common denominator in these studies is the individual's ability to exercise agency in digital domains. At the same time, it becomes clear that current legislative processes are riddled with misconceptions about user empowerment. When given the opportunity to reflect, people easily demonstrate a basic understanding of what is best for them. This demonstrates that transparency should be observed from the user's perspective rather than the platform's. It is up to policymakers to enact a regime in which platforms provide sufficient, adequate, or meaningful situations for human reflection. In other words, the foundational pillar of meaningful transparency should be the 'boosting' of individual capacities.

Boosting is a class of behavioural interventions that aim to foster competencies instead of steering behaviour. In the context of transparency, boosting means fostering the *understanding* of information instead of the mere *provision* of information.[45] It empowers users to translate information into actions aligned with their preferences. Accordingly, platform regulation should change the objective of transparency for transparency's sake to *measurable understanding*. After considering the shortcomings of current transparency regimes and anticipating the development of interoperable technologies, a future-proof international framework should consider implementing meaningful transparency requirements at the platform level, resulting in individual empowerment.

To incorporate the above findings into UNESCO's guidance for regulating platforms, we suggest the following amendments to the document:

Para 25: Transparency:

Add Para 25.7 as follows:

25.7 To achieve the above objectives, platforms should conceptualise 'transparency' as 'meaningful transparency'. Instead of equating transparency with the provision of legal texts, 'meaningful transparency' should be understood as providing users with the information they desire and require to make informed decisions. The effectiveness of platforms' transparency mechanisms should be independently evaluated through qualitative and empirical quantitative assessments to determine whether the information provided for meaningful transparency has served its purpose. Reports should be made available to users on a regular basis. Failure to provide 'meaningful transparency' to users should be integrated in the risk assessment as proposed under para. 33.2 and lead to concrete and measurable mitigation measures.

Summary of Main Discussion Points

- Regulatory regimes around the world have faced significant challenges in regulating technology due to the complex nature of technological development and processes. Regulators have established regulatory regimes that center around the concept of transparency, such as the European Union's Digital Services Act (DSA) and General Data Protection Regulation (GDPR).
- Despite the heavy reliance on the principle of transparency, the term is often vaguely defined and results in large bodies of technical or legalistic text. This has led to the emergence of the transparency fallacy (transparency in name only) as a characteristic of European digital legal regimes.

[45] Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. Perspectives on Psychological Science, 12(6), 973–986. https://doi.org/10.1177/1745691617702496
For examples, see https://scienceofboosting.org.

- Critics argue that transparency techniques may not yield the intended benefits as they rely on consumers to process and act rationally on the information provided. Transparency requirements in the form of privacy policies and consent notices have led to "consent fatigue" and "rational ignorance" among users, rather than empowering them to make informed decisions about their digital engagement.
- The term 'dark patterns' is commonly used by the web collective to describe a deceptive user interface that exploits users into doing something they would not normally do. It is a coercive and manipulative design technique used by web designers. Complex deterministic and stochastic (non-deterministic) dark patterns cannot be identified, even by an expert. They require inside information on the inner workings of the design interface and the system architecture. These patterns compromise legal requirements such as consent and privacy-by-design, as well as legal principles such as fairness and transparency.
- Algorithmic transparency, in its present state, is based on the fallacy of assuming a direct link between visibility and understanding. The combination of deceptive design and AI-powered personalization of decision environments suggests that we are in a new era of manipulative techniques used to benefit platforms at the expense of user autonomy.
- A more dynamic approach to platform regulation is needed. Keeping the status quo in place will only ensure the continued flow of disinformation and harmful content, despite evidence that users would judge content differently if provided with a choice to do so.
- Transparency-centred regimes should focus on the clear implementation of meaningful transparency and measurable or effective transparency. Meaningful transparency describes the disclosure process of the most relevant, targeted information that would benefit the intended recipient.
- Measurable and effective transparency describes the process through which platforms test and evaluate the efficiency of the tools used to provide users with meaningful and transparent information.
- The foundational pillar of meaningful transparency should be the 'boosting' of individual capacities. Boosting is a class of behavioural interventions that aim to foster competencies instead of steering behaviour. In the context of transparency, boosting means fostering the understanding of information instead of the mere provision of information.

21.3 Platforms empower users to use digital services in a self-determined and empowering manner, including assessing the quality of the information received

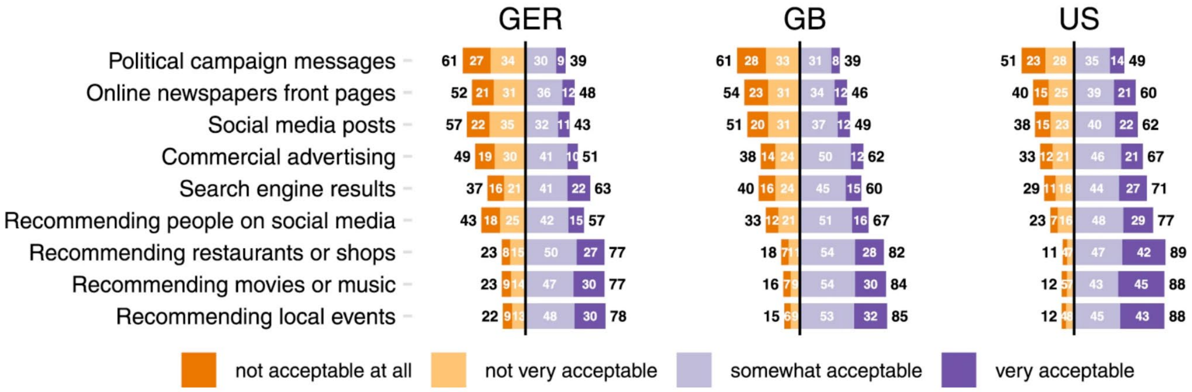
Interoperability of Personalised Settings

Interoperability refers to the ability of information systems to exchange data and facilitate the sharing of information.[46] This is particularly important for the exchange of data across platforms, as it allows for competition from privacy-friendly competitors and the ability for users to move their data to providers with more favourable privacy settings. Interoperability is a multi-layered concept, with the third layer focusing on standard underlying models and the codification of data through the use of standardized definitions and vocabulary. The fourth layer includes considerations such as governance, policy, and social and legal factors to facilitate secure and seamless communication and use of data.

Privacy advocates have called for default browser settings to be set to the most privacy-protective by default. However, this approach does not consider individual user preferences and autonomy. To truly empower users, a default, generalized setting of user preferences in conjunction with interoperability would allow users to engage their privacy settings, cookie consent, and content moderation settings once, preferably on their own devices. These settings could then be transferred to other providers through interoperability obligations. Research has shown that users do not want to repeatedly update their settings or be subject to a single "take-it-or-leave-it" approach to personalization and recommendation systems. Interoperability of personalized settings would therefore protect user autonomy and replace the imposition of a content moderation policy that primarily benefits the economic interests of a platform.

If user empowerment is a stated objective of platform regulation, a default, generalized setting of user preferences combined with interoperability would allow users to engage their privacy settings, cookie consent, and content moderation settings once, preferably on their own devices. These settings could be transferred to other providers due to interoperability obligations. It is axiomatic that users do not want to be bothered with repeated requests to update their settings. As Kozyreva et al.'s study show (**Figure One**), people are far more open to the use of their personal data for recommending local events, movie and music, and search engine results than they are for personalizing political campaign messaging.

Figure One[47]:



[46] https://edps.europa.eu/data-protection/our-work/subjects/interoperability_en

[47] Adapted from Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021). Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the US. Humanities & Social Sciences Communications, 8(117). https://doi.org/10.1057/s41599-021-00787-w

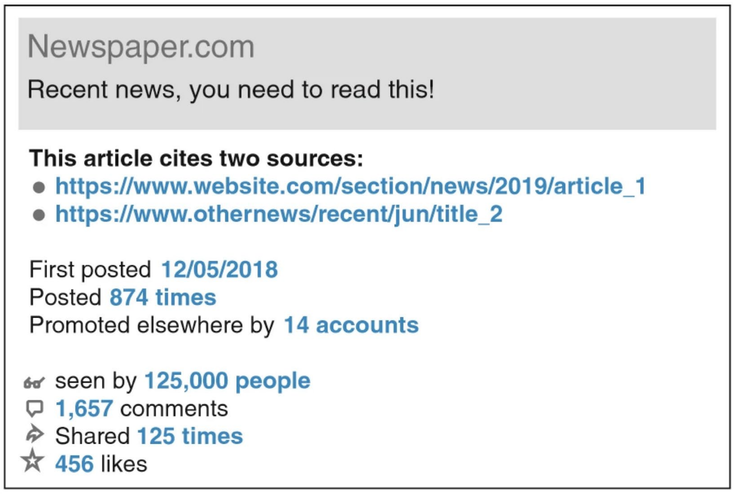
A ‘Right to Cues’

The current design practices of digital platforms often prioritise a "frictionless" user experience over transparency, resulting in a lack of information about the background and context of the content presented to users. The "feed" structure, for example, presents all posts in a similar format, with little distinction between advertisements, personal opinions, and journalistic content. This structure aims to present users with as much content as possible, often without providing relevant information about the context or history of the posts. This lack of transparency contributes to the knowledge asymmetry between platforms and users, who are often treated as passive consumers rather than informed actors.

To address this issue, we propose implementing a "right to cues" that would provide users with standardized information, such as context and background, to enable informed judgement about the relevance and quality of a piece of content.[48] This can be achieved through various methods, such as providing relevant information about a post or article without rating the article itself or making links to other sources in an online article readily available.[49] Additionally, transparency in the process of sharing and curation algorithms that reveal factors that explain why a particular post was selected as more relevant than others can also be implemented to develop digital literacies.

The "right to cues" should include a standardized set of information (possibly with iconography) that provides brief information about the context and background of the information (For an example, see Figure Two). This approach borrows from the legal obligations of the food industry to provide nutritional information on packaging. Notices must provide relevant information about a post or article without rating it (avoiding human rights implications). This gives users the means to infer the quality of the content without judgement.

Figure Two[50]:



Newspaper.com

Recent news, you need to read this!

This article cites two sources:

- https://www.website.com/section/news/2019/article_1
- https://www.othernews/recent/jun/title_2

First posted 12/05/2018
Posted 874 times
Promoted elsewhere by 14 accounts

seen by 125,000 people
1,657 comments
Shared 125 times
☆ 456 likes

A recent example of this is Twitter's addition of the number of impressions of a tweet to existing engagement metrics (Likes, Shares, Replies), contextualizing these statistics with a natural baseline rate of how many people saw the post. This is just one example of a wide range of possibilities for relevant cues that should be made easily accessible to users so that they have a chance to act as empowered consumers in the first place. For example, making links to other sources in an online article readily available, or providing transparency in the process of sharing and curation algorithms that reveal factors that explain why a particular post was selected as more relevant than others, would be the first step in enabling the development of digital literacies.

[48] similar to nutrition labels in the food industry (see Figure 2 for a fictional example)

[49] Wineburg, S., & McGrew, S. (2017). Lateral reading: Reading less and learning more when evaluating digital information

[50] Adapted from Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R., & Hertwig, R. (2020). How behavioural sciences can promote truth, autonomy and democratic discourse online. Nature human behaviour, 4(11), 1102-1109.

In summary, in addition to meaningful transparency and the prevention of deceptive design, providing users with informative cues about the content they see is crucial for truly empowered users and balancing the asymmetry between platforms and users. This can be achieved through various methods, such as providing context and background information or linking to other sources readily available.

Summary of Main Discussion Points

- Interoperability refers to the ability of information systems to exchange data and facilitate the sharing of information, particularly across platforms, which allows for competition from privacy-friendly competitors and the ability for users to move their data to providers with more favourable privacy settings.
- To empower users, a default, generalized setting of user preferences in conjunction with interoperability would allow users to engage their privacy settings, cookie consent, and content moderation settings once, preferably on their own devices. These settings could then be transferred to other providers through interoperability obligations. This approach considers individual user preferences and autonomy, and research has shown that users do not want to repeatedly update their settings or be subject to a single "take-it-or-leave-it" approach to personalization and recommendation systems.
- The current design practices of digital platforms often prioritize a "frictionless" user experience over transparency, leading to a lack of information about the background and context of the content presented to users.
- To address this issue, we propose implementing a "right to cues" that would provide users with standardized information, such as context and background, to enable informed judgement about the relevance and quality of a piece of content. This can be achieved through various methods, such as providing relevant information about a post or article without rating the article itself or making links to other sources in an online article readily available.

21.4 Platforms are accountable to users, the public, and regulators in implementing terms of service and content policies, including giving users rights of redress against content-related decisions.

The accountability of online platforms means that the individual interests of the company can only be pursued if they not only do not harm legal goods of social relevance but take important and effective steps to ensure meaningful user empowerment. Platform Regulation 2.0 should, therefore, emphasise the empirical measurement of effectiveness subject to audit over meaningless buzzwords or nominal transparency, shift the burden of proof about the empowerment of users onto platforms and ensure users are actively protected from dark patterns and other forms of manipulative design. That could mean that platforms need to empirically show that certain benchmarks of measurable understanding of their updated policies or designs are reached (e.g., show that more than 50% of a random user sample answered a few simple questions about a new policy correctly). Best practices for healthy and empowering user interfaces should be standardised in the best interests of the user rather than for the economic benefit of the platform. In practice, this would mean a shift in the ethos of platforms' business models. Instead of subjecting human rights law to its profits, corporate objectives need to first find validity in fundamental and human rights frameworks.

To incorporate the above findings into UNESCO's guidance for regulating platforms, we suggest the following amendments to the document:

Para 48: Power

Add para 48:

In-scope digital platform services will be required to report regularly how they are achieving the goals. Platforms should hereby bear the burden of proof and demonstrate the steps taken to establish that a transparency-oriented operation was fruitful in generating the user's active participation. Regulators may commission off-cycle reports if there are exigent circumstances, such as a sudden information crisis (such as that brought about by the COVID-19 pandemic) or a specific event which creates vulnerabilities (e.g.elections, protests, etc.).

Conclusion:

The United Nations, through the Windhoek Declaration, recognizes the importance of information as a fundamental public good that is vital for the advancement of human rights.[51] As such, the right to access information and the right to free expression are critical for ensuring principles such as non-discrimination and gender equality is upheld. However, the violation of these rights has been exacerbated by the nature of human interactions on digital platforms. Through their content moderation policies, these platforms have become gatekeepers of information, limiting individuals' control over the information they receive. To carry out this gatekeeping role, platforms utilize opaque algorithmic systems to generate personalised but non-customizable content moderation. In light of this, this report explores the ways in which platforms may cause distortions to fundamental rights in terms of transparency, user empowerment, and accountability.

Transparency requirements in regional regulations were intended to improve regulatory control over platforms and provide users with detailed information on algorithmic processes. In practice, transparency requirements resulted in voluminous and frequently incomprehensible legalistic text. Despite an increasing body of evidence that transparency, as a legal concept, remains nebulous and

[51] UNESCO, Windhoek + 30 Declaration: information as a public good, World Press Freedom Day 2021, (2021)

abstract for users, there is little evidence that the principle is helping people make better and more informed decisions. These lengthy explanations are unacceptable within an international regime aiming to protect fundamental human rights. Instead, the legal concept of transparency should be interpreted through the lens of meaningful and effective transparency.

This report examines various studies that focus on the individual's ability to engage in autonomous activity online. The findings indicate that providing appropriate and relevant information can encourage active participation in the digital environment. Research suggests that users are willing to remove false and harmful content, as well as punish those who repeatedly post such content.[52] The report suggests that empowering users with the ability to personalize their own content moderation policies through interoperability requirements ensures that users see the content they desire rather than content that aligns with a business's financial growth model. The report also highlights the importance of dynamic personalization and responsive regulation that is informed by evidence-led insights. For example, the report suggests that instead of binary choices through cookie selection, increasing autonomy and control over the type of algorithmic personalization better reflects users' actual preferences.[53]

Meaningful transparency refers to the ability of individuals to make informed decisions by providing relevant information. This information should prioritize factors deemed important by users rather than solely what platforms consider relevant. Various techniques have been proposed to enhance meaningful accountability within legislative and platform systems, including the creation of interoperable personalized settings and the recognition of the right to information (cues). These solutions allow regulators, researchers, and platforms to conduct empirical evaluations to assess the effectiveness of transparent prompts, which is known as measurable and effective transparency.

Effective and measurable transparency is achieved through the implementation of legislation requiring platforms to conduct both qualitative and quantitative assessments to determine the effectiveness of the information provided in the name of transparency. These assessments, which measure the user's reaction to various transparency-oriented prompts, ensure that platforms uphold and promote the fundamental rights of their users. These evaluations also ensure accountability for users and regulators, as they provide active and autonomous participation for users and a system of accountability for regulators.

International accountability systems should hold platforms accountable for providing evidence of their efforts to establish transparency-oriented operations, such as personalized interoperable settings or cues, that effectively encourage user engagement. By enforcing accountability on platforms, regulations focused on meaningful transparency can protect fundamental human rights, including access to information and freedom of expression. The United Nations has a crucial role in promoting legislation that reflects principles of platform accountability, measurable transparency, and user empowerment. When implemented effectively through international regulation, these principles benefit all parties involved, resulting in a mutually beneficial international environment.

[52] Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., & Reifler, J. (in press). Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences of the United States of America*. Preprint: https://doi.org/10.31234/osf.io/2pc3a

[53] Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021). Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the US. Humanities & Social Sciences Communications, 8(117). https://doi.org/10.1057/s41599-021-00787-w