ACCOUNTABILITY REPORT 2.0

An independent evaluation of online trust and safety practice.

THE INTERNET COMMISSION

10th March 2022

Authors:
Jonny Shipp, Anno Mitchell, Ioanna Noula, Patrick Grady

Acknowledgements:
Juraj Kosturik, Emma Hoksbergen, Aebha Curtis, Jacob Bendicksen, Ruairí Harrison

THE INTERNET COMMISSION

Foreword

Kersti Kaljulaid
Former President of Estonia

Introducing the Internet Commission's Accountability Report 2.0

What options are there for companies who find one day that their impact on human societies is far bigger than the wildest dreams (or fears) of the founding fathers? How to preserve the good and eliminate the bad of what your sector appears to be doing to humankind? How to demonstrate corporate responsibility and therefore avoid potential overregulation?

This is the motivation for the Internet Commission's Accountability Report 2.0. The first edition was a methodological trial. With the second, the Internet Commission has made significant progress. Evaluations of the participating enterprises are structured in an easy to read and understandable way. Therefore, it achieves at least two objectives. First, for participating enterprises, an honest health check and way forward is established. Second, for all other companies interested in similar assessments, the report provides inspiration and encouragement.

Ideally, in a world transformed by digital technologies, an algorithm would be established to report on how a technology company manages its societal risks. But to this day, it remains only a dream that each and every technology becomes capable of self-explanation, self-monitoring and self-reporting of anomalies that are potentially dangerous to humankind.

Luckily, people working in tech recognise that in order to not be regulated out of the innovative and transformative potential of their technologies and their creative applications, certain standards need to be established and followed. I wholeheartedly support the Internet Commission in its approach and I am eagerly looking forward to mainstreaming the methodology described in the report so that all interested parties can feel we all understand the subject matter in a similar way, and are able to describe, discuss and if necessary, negotiate the sensitive topic of the societal impact of technologies and tech companies globally. This is the way to go for a free world, to protect privacy, establish content presentation neutrality, guarantee transparency and accountability to regulators, politicians and users.

Common ground must be found to make sure that the free world of democracies does not lag behind other societies that do not have to answer transparency and accountability questions. A permissive and protective environment is needed, and the Internet Commission is taking important steps towards exploiting the capabilities of the tech sector for the common good.

EVERY WOMAN
EVERY CHILD
GLOBAL ADVOCATE OF THE UNITED
NATIONS SECRETARY-GENERAL

In 2016 Kersti Kaljulaid was elected President of the Republic of Estonia. During her presidency, she has been a vocal advocate of human rights, rule of law, freedom of speech and democracy. Previously she had been serving as a Member of the European Court of Auditors, advising Prime Minister Mart Laar and holding different top-level positions in energy, investment banking and telecom sector. President Kaljulaid has become the first Estonian to be featured in the Forbes World's 100 Most Powerful Women. In 2021, she was appointed the first Global Advocate of the United Nations Secretary-General for Every Woman Every Child.

Preface

Christopher Hodges
University of Oxford

Delivering Trust, Evidence and Ethics

Humans have a built-in ability to distinguish between right and wrong, and we use this to differentiate between those we can trust and those we distrust, based on the available evidence. This mechanism of ethical evaluation forms the basis of our ability to cooperate in groups—as any good corporate manager will know. Markets and regulatory systems have evolved particular ways of building trust and relevant evidence on which to base it. An important stream of evidence comes from the legal system. Examples are contracts that contain specific obligations and can be enforced through the courts, standards, auditing and accreditation systems, or regulatory requirements and licences that can be enforced by authorities or through courts. The disruptive forces unleashed by globalisation and digital technology mean that we are having to revolutionise the types of evidence that we want to be able to rely on and the mechanisms for producing it. The opportunities for relying on extensive data and its aggregation, and interrogation through Artificial Intelligence (AI), are enormous. So we debate whether we can trust new forms of data, whether its sources, controllers, systems and AI processes are reliable. Can these be trusted, and on what basis? Should we standardise and regulate them and how?

Balázs Bodó has argued that the new class of private trust producers —online reputation management services, distributed ledgers, and AI-based predictive systems—commodifies trust and causes real issues of trustworthiness in relation to the substance of the information being spread. We know that we cannot single-mindedly pursue personal goals to the exclusion of others' interests. Maximising shareholder value has crumbled in the face of systemic disasters such as the possible extinction of life on this planet, the global financial crisis, pandemics, and possible further major wars. What is needed to defeat these real threats is cooperation, which involves strengthening trust and common values and goals. It means looking hard at how we are going to achieve the UN's Sustainable Development Goals and about business genuinely adopting social and environmental purposes, making stakeholder value real.

In addition, we now know much more than we did about how and why humans achieve goals and also make (sometimes huge) mistakes. We know that humans can be diverted by lack of time or attention, by focusing on certain (sales or financial) targets, or unethical objectives (e.g. criminal gangs). We also know that people are strongly influenced by what others around us are saying and doing—so social (i.e. organisational) culture is important. We know that people will not share information that may be embarrassing unless they feel psychologically safe. We know that 'hard enforcement' tools used on people who think that they are trying to do the right thing are not only ineffective in delivering 'compliance' but can also be counterproductive. On the other hand, we know that supporting people's intrinsic motivation—their senses of competence, autonomy and relatedness—are effective in getting the best out of people.

Putting all this together, good leaders, managers and regulators all strive to create empowered, confident, and ethical individuals and cultures. It is these elements that differentiate societies, nations and institutions around the globe and increasingly in the commercial world. Behaviour of any sort—and certainly commercial and social behaviour—will not drive successful outcomes unless it is fair and transparent, and hence trusted. The real challenges we face at the moment are how to devise reliable mechanisms for producing evidence that we think is relevant, comprehensive enough, based on common ethical values. It is these elements that deserve experimentation and open debate. We have had an extensive debate over ethical values and have produced many (even a confusing number of) statements about them. But now we have to be practical. What should our outcomes be—and not be? What mechanisms (technical, AI and human) are available and reliable?

The evidence of behavioural science points to the advantages of cooperation and co-creation in making rules or codes and evaluating systems and outcomes. Such open and cooperative modes may be challenging for governments, MPs, regulators, companies and others. But such systems do work and support innovation well.

It is for these reasons that the Internet Commission is playing an invaluable role in contributing to the development of a system of trust in the digital world by discussing what companies do and how they seek to achieve ethical outcomes and control against harms. This contributes to an evolving picture of what good looks like, how it is achieved, how well it is achieved, and how it needs to be changed. A toolbox of practices is evolving that can be adopted more widely. There is an attempt to balance transparency and commercial confidentiality, opening practices and outcomes to scrutiny by a range of disinterested experts and civil society representatives. This is an iterative, evolutionary process. The annual rounds point up issues to focus on as next steps as each cycle progresses. One interesting current challenge is how we can crystallise metrics on outcomes (rather than on outputs or just processes) that demonstrate progress towards achievement of desirable goals and good outcomes, and reduce incidence and risk of undesired outcomes. Another conundrum is how to measure organisational culture and especially ethical culture. In any event, these endeavours to build trust, and demonstrate commitment and improvement, are not just worthwhile but essential.

> *Professor Christopher Hodges OBE MA PhD FSALS FRSA is Emeritus Professor of Justice Systems, University of Oxford; Supernumerary Fellow of Wolfson College, Oxford; Co-Founder, International Network for Delivery of Regulation (INDR).*

Independence

The Internet Commission's scrutiny group consists of five external experts from civil society, academia and industry, who have no other association with, or interest in, the participating organisations. It provides independent scrutiny of this accountability report.

> *"The scrutiny group reviewed this report and the case files upon which it was prepared during November 2021. It discussed these with the report's authors and had the opportunity to question them and obtain any necessary clarifications.*
>
> *Based on our review of the evidence provided by the participating organisations, we believe that this report offers an independent view of how organisational cultures, systems and processes align to support corporate digital responsibility, with a particular focus on internet safety, freedom of speech and the ways in which decisions are made in relation to content, contact and conduct online.*
>
> *The conclusions appear to be consistent with the detailed evidence gathered. We do not believe that these conclusions have been shaped to serve any particular individual, commercial or other interest.*
>
> *We did not conduct any form of audit or external verification: our review was based only on this report, the case files and evidence that was used to prepare it, and the original information which was provided by the participating organisations themselves."*

Brian O'Neill (Chair)
Emeritus Professor at Technological University Dublin

Paul Adamson
Chairman of Forum Europe and founder and editor of Encompass

Bojana Bellamy
President of Hunton Andrews Kurth LLP's Centre for Information Policy Leadership

Sheena Horgan
CEO of Drinkaware Ireland

Claire Milne
Independent ICT policy consultant

Contents

Executive Summary

This report presents the findings and analysis of the Internet Commission's second reporting cycle. It looks at how organisational cultures, systems and processes shape online experiences and comprise ethical business practice.

We focus on internet safety, freedom of speech and the ways in which decisions are made about content, contact and conduct online[1]. We found that leading organisations (i) value the voice of their users, (ii) demonstrate coordinated oversight to anticipate negative impacts, and (iii) use innovation to balance safety and freedom online.

Our Evaluation Framework has evolved based on learning from the first reporting cycle, important policy developments and feedback from experts. The resulting second framework comprises four sections: Organisation, people and governance; Content moderation; Automation; and Safety. Based on an in-depth evaluation process, we use an organisational maturity model to scale the maturity of organisations and their trust and safety practices. We identify common challenges, and observed patterns, before detailing a total of 46 practices.

Our analysis explains how these practices determine the impact organisations have on individuals and society. The effects of some practices are external to organisations: reinforcing user agency, raising expectations and standards of individual behaviour, and shaping positive online experiences. The effects of others are internal, building and integrating an organisation's capacity for ethical business practice and leadership across its various business operations.

We conclude by discussing three emergent themes. First, the importance of engaging users in the strategic shaping of policy. Second, the need for deliberate oversight to take responsibility for social impact. Third, the role of innovation in delivering both safety and freedom in the online and offline world.

We look forward to discussing our work and building on our findings and analysis.

[1] The '3Cs' is a well-established framework for understanding categories of risk. The framework has recently been updated: https://core-evidence.eu/updating-the-4cs-of-online-risk/.

1. Introduction

The Internet Commission's accountability report is premised on the idea that corporate accountability can contribute to healthier digital environments, trusted by citizens and consumers, and underpinned by balanced and informed regulation.

The Internet Commission has established an innovative, evidence-based approach to evaluating the ways in which organisations address the consequences of digitalisation. Our insights and recommendations offer valuable insights for policymakers and seek to contribute to wider, multi-stakeholder deliberations. For the organisations involved in reporting, our process is a tool with which to manage risk, improve practice and demonstrate accountability.

This second accountability report was developed against the backdrop of significant digital policy developments, which seek to improve the social impact of digital services. In December 2020, the European Union presented its proposal for a Digital Services Act (DSA)[2] and, in May 2021, the UK government published the Online Safety Bill (OSB)[3].

The DSA aims to put pressure on organisations that operate digital services to (a) protect the wellbeing and fundamental rights of citizens online; (b) set clear transparency and accountability frameworks, and; (c) attend to systemic risks posed by the misuse of their services in relation to the dissemination of disinformation and illegal content. The OSB will introduce new duties of care and codes of practice, requiring organisations that operate digital services to act against illegal and harmful content. Critics of both proposals raise concerns about risks to freedom of expression, vague wording, and a failure to account for the breadth of online services.

The Internet Commission works with organisations that wish to move beyond compliance approaches to corporate digital responsibility (CDR). Drawing on state-of-the-art, holistic approaches to corporate responsibility including corporate purpose[4] and ethical business practice[5] from sustainability, legal and business literature, we have adopted the following definition of CDR:

The set of shared values and norms guiding an organisation's operations with respect to four main processes related to digital technology and data: (i) the creation of technology and data capture, (ii) operation and decision making, (iii) inspection and impact assessment and (iv) refinement of technology and data[6]*.*

[2] "The Digital Services Act Package"(2021) Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.

[3] "Draft Online Safety Bill" (2021) GOV.UK. https://www.gov.uk/government/publications/draft-online-safety-bill.

[4] Colin Mayer, Prosperity: Better Business Makes the Greater Good, Oxford, Oxford University Press (2018).

[5] Christopher J. S. Hodges, Ruth N. Steinholtz, Ethical Business Practice and Regulation: A Behavioural and Values-Based Approach to Compliance and Enforcement, Oxford, Hart Publishing and London, Bloomsbury Publishing (2017).

[6] Lara Lobschat, Benjamin Muellerb, Felix Eggersd, Laura Brandimartee, Sarah Diefenbachf, Mirja Kroschkea, Jochen Wirtz, "Corporate Digital Responsibility", Journal of Business Research 122 (2021) 875-888, https://doi.org/10.1016/j.jbusres.2019.10.006

Key takeaways

1. Value the voice of users

To improve the design and development of their services, and to better protect users from harm, organisations consult a range of stakeholders: industry specialists, policymakers, and experts from civil society. We found that, as well as engaging external stakeholders, leading organisations recognise the value of the user voice. Through regular consultations, and by building forums for feedback, organisations draw valuable insights from the experiences and expertise of their users.

Twitch, for example, invites content creators to take part in its advisory council. Involving users in this way democratises the process of policy and product development, empowers users, and builds trust between the service and user.

2. Coordinate oversight to anticipate negative impact

Organisations should take responsibility for the expected impact of their services. This starts with ensuring that the product and operations of their services are designed to anticipate potential impacts. Tinder, for example, configures its moderation technology to learn of, and respond to, new kinds of harm. Proactive and coordinated oversight of standards demonstrates ethical intent.

Maintaining oversight is especially important when deploying automated tools, which can misfire, unfairly punish users and amplify biases online. Recognising these risks, Twitch chooses to keep humans in the loop when moderating content.

3. Balance safety and freedom through innovation

Organisations that offer digital services often face the challenge of balancing safety and freedom online. In some cases, the nature of the platform will dictate what is most appropriate: age assurance monitoring is more appropriate in services that must assure the safety of younger users; moderating speech, to mitigate the risk of harm to populations, is less appropriate in closed, one-to-one communications.

In other cases, however, innovations in technology and processes can remove such trade-offs between safety and freedom. The parent company of Meetic and Tinder, for example, has developed tools that encourage potential perpetrators to rethink their harmful message before sending, and potential victims to submit reports of harm. In the context of online dating, these tools encourage a safer environment whilst, at the same time, maintaining a user's freedom to decide.

Considerable commitment, investment and iteration are necessary to find the right solutions, which can vary significantly between different organisations and services.

2. Methodology

The Internet Commission's Evaluation Framework and Evaluation Process together offer a holistic annual reporting cycle, which crosses existing policy and sectoral silos and contributes to the development of ethical business cultures. Asking how and why things are done, and not yet presuming particular transparency metrics, our evaluation aims to produce structured and consequential insights into how cultures and processes shape organisational outcomes.

This approach goes beyond compliance thinking, seeking to shed light on the way decisions are made and the extent to which organisations consciously and proactively act in the interests of society. It recognises the interdependence of shareholder and stakeholder value.

Evaluation Framework

The first version of the Internet Commission's Evaluation Framework for Content Moderation was published in December 2019 following 18 months of consultation with experts, regulators, policymakers, public interest groups and companies in the UK, Europe and Australia. It incorporated the 2018 Santa Clara Principles on Transparency and Accountability in Content Moderation[7], and reflected the growing importance of AI and Machine Learning (ML) and the balance of safety, security, privacy and freedom of expression.

Our Evaluation Framework was updated in 2021 to incorporate learnings from the first reporting cycle. First, we took advantage of the rich knowledge base generated from the responses to our framework questions, supplementary questions and interviews. By reviewing these responses, we uncovered gaps to close and new themes to consider. Second, we updated the content and structure of the framework, considering important policy developments, emerging issues and key indicators being used in adjacent areas[8]. Third, we shared the framework with international experts across sectors who advised on priority areas for investigation and the structure of the framework.

The resulting second edition[9] looks at how organisational cultures, systems and processes align to support corporate digital responsibility. It has a particular focus on internet safety, freedom of speech

[7] "The Santa Clara Principles, On Transparency and Accountability in Content Moderation" (2017) https://santaclaraprinciples.org/

[8] 2020 Ranking Digital Rights Corporate Accountability Index" (2020) Ranking Digital Rights. https://rankingdigitalrights.org/index2020/explore-indicators; "Voluntary Principles To Counter Online Child Sexual Exploitation And Abuse". 2020. Weprotect.Org. https://www.weprotect.org/wp-content/uploads/11-Voluntary-principles-detailed.pdf.; "Introduction To The Age Appropriate Design Code"(2021) Ico.Org.Uk. https://ico.org.uk/for-organisations/guide-to-data-protection/ico-codes-of-practice/age-appropriate-design-code/; Livingstone, Sonia, Eva Lievens, and John Carr (2021) "Handbook For Policy Makers On The Rights Of The Child In The Digital Environment". https://rm.coe.int/publication-it-handbook-for-policy-makers-final-eng/1680a069f8.

[9] "Evaluation Framework For Digital Responsibility" (2021)

and the ways in which decisions are made about content, contact and conduct online. It comprises the following four sections:

1. *Organisation, people and governance:* about the organisation's scope and purpose, the people concerned and its governance.
2. *Content moderation:* asks how harmful and illegal contact, content, or conduct is discovered and acted upon.
3. *Automation:* asks how intelligent systems are used to promote or moderate online content.
4. *Safety:* asks what measures are in place to protect people's health and well-being.

Evaluation Process

Our Evaluation Process was developed in 2019 and first implemented in 2020. It has three phases, designed to enable organisations to review their practices, share knowledge with peers and participate in an independent public report, whilst maintaining legitimate commercial confidentiality.

1. Review

Data is collected using a questionnaire based on the Evaluation Framework. Based on a review of this data, interviews with front-line staff and senior management are prepared and conducted. Individual, confidential case studies are then drafted and reviewed with the organisation concerned. Amendments are made as necessary and a final version is agreed upon. For the organisations, these confidential case studies provide a tool for reflection, and further development, of their processes and practices. A second, "redacted" version of this case study is then prepared and agreed upon, removing any information and details that the organisation does not want to share with its peers.

2. Explore

When all the redacted cases are completed they are combined into a draft of this report. This draft is used as preparation and pre-read for a knowledge-sharing workshop. The workshop brings together representatives of participating organisations to discuss best practices, key challenges and innovation in trust and safety. Collaborating in this way provides a unique forum for the organisation and has revealed the need for these communities to connect and gain insight from each other. As well as helping to shape the final public report and test its conclusions, this process contributes to industry-led capacity building that advances digital responsibility.

3. Engage

A draft public accountability report is prepared and reviewed by participating organisations. This is also reviewed by a "scrutiny group" comprising experts from businesses, civil society and academia.

Members are given access to confidential files and data gathered in order to discuss the draft report in detail with its authors. Based on this, amendments are negotiated with the participating organisations and the report is launched as a stimulus for an ongoing multi-stakeholder dialogue.

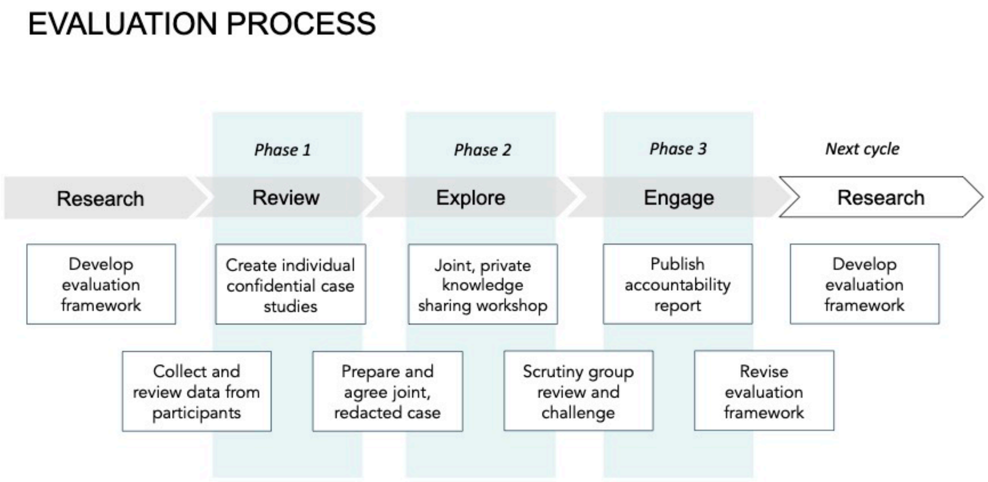

EVALUATION PROCESS
Phase 1
Phase 2
Phase 3
Next cycle
Research
Review
Explore
Engage
Research
Develop evaluation framework
Create individual confidential case studies
Joint, private knowledge sharing workshop
Publish accountability report
Develop evaluation framework
Collect and review data from participants
Prepare and agree joint, redacted case
Scrutiny group review and challenge
Revise evaluation framework

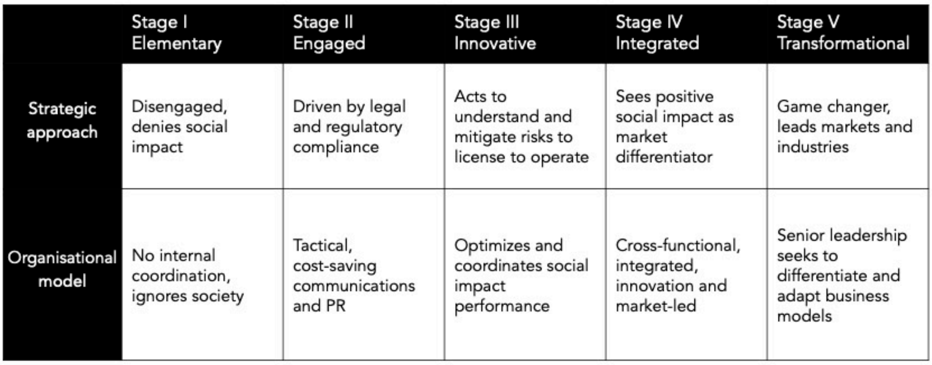
Figure 1

Maturity Model

Drawing on literature from the field of Corporate Social Responsibility[10], we used an organisational maturity model (Figure 2) to evaluate an organisation's practices and its wider social impact. We asked, what does this practice tell us about the organisation's strategic approach and model for digital responsibility? We used a five-stage scale to evaluate the maturity of participating organisations by exploring and testing the congruence of observed practices and the organisation's digital responsibility goals.

[10] Głuszek, Ewa (2018) "Dimensions And Stages Of The CSR Maturity". Prace Naukowe Uniwersytetu Ekonomicznego We Wrocławiu, no. 520: 64-80. doi:10.15611/pn.2018.520.06.

MATURITY MODEL



	Stage I Elementary	Stage II Engaged	Stage III Innovative	Stage IV Integrated	Stage V Transformational
Strategic approach	Disengaged, denies social impact	Driven by legal and regulatory compliance	Acts to understand and mitigate risks to license to operate	Sees positive social impact as market differentiator	Game changer, leads markets and industries
Organisational model	No internal coordination, ignores society	Tactical, cost-saving communications and PR	Optimizes and coordinates social impact performance	Cross-functional, integrated, innovation and market-led	Senior leadership seeks to differentiate and adapt business models

Adapted from: Głuszek, Ewa. (2018). CSR Maturity Model – Theoretical Framework. Journal of Corporate Responsibility and Leadership. 4. 25. 10.12775/JCRL.2017.015.

Figure 2

Methodological contribution

The Internet Commission is building a position as a trusted broker within a new regulatory system. It holds a growing body of data about organisational practices related to digital responsibility. With this expertise, it aims to ask the right questions, provide reliable evidence, generate new insights for stakeholders, and help organisations navigate different national and international standards and regulatory requirements. This approach moves away from the model in which organisations set their own questions for transparency reporting.

Although data is provided by the participating organisations themselves, this analysis is supported by depth interviews and our growing capacity for comparative analysis. The Evaluation Process enables key practices to be identified and scrutinised by independent experts and civil society representatives, in a manner that seeks to amplify success more than to call out failure, steering away from the “name and shame” approach that discourages the development of socially accountable organisational cultures.

This approach has allowed the Internet Commission to achieve unique access to digital organisations and create a space where good practices, challenges and other confidential information can be discussed. Knowledge sharing workshops allow new thinking to be explored, substantiating our analysis, and creating opportunities for collaboration and improvement. Looking ahead, this approach may provide an opportunity and foundation for the eventual co-creation of appropriate quantitative, outcome-focused metrics.

3. Findings

This chapter reviews the state of the art, considering what the identified practices reveal about the culture of participating organisations and how these operational choices affect online experiences and wider society. Our goal is to support the development of best practices, to enable shared insights and to help organisations demonstrate accountability, rather than rank or shame companies. We have therefore agreed with some organisations to include their less developed practices in an anonymised form.

This chapter presents 46 practices identified during the Internet Commission's 2020 and 2021 accountability cycles. 23 were newly identified in 2021. Of the 23 that were first identified in 2020, 5 have been updated. These practices are organised into the four sections of the Evaluation Framework:

1. Organisation, people and governance
9 practices, 5 newly identified in 2021;
4 were first identified in 2020, 1 of which was updated in 2021

2. Content moderation
16 practices, 7 newly identified in 2021;
9 were first identified in 2020, 1 of which was updated in 2021

3. Automation
7 practices, 3 newly identified in 2021;
4 were first identified in 2020, 1 of which was updated in 2021

4. Safety
14 practices, 8 newly identified in 2021;
6 were first identified in 2020, 2 of which were updated in 2021

Reporting organisations (2020 & 2021):

The **BBC** is the UK's national broadcaster, the world's oldest and largest national broadcaster, with a website and news service reaching an international audience.

Meetic is a closed, peer-to-peer platform offering an online dating service with active members across 15 European countries in 11 languages: Danish, Dutch, English, Finnish, French, German, Italian, Norwegian, Portuguese, Spanish and Swedish.

PopJam is a curated content-sharing app for children aged 7-12 to create and share art and photos, participate in quizzes and play games.

Sony Interactive Entertainment (SIE) provides a global games platform (PlayStation) and is also a games publisher (PlayStation Studios). PlayStation offers fast-paced social experiences in and out of gameplay.

Tinder is an online dating platform. It operates a 'freemium' model, where users can enjoy core features for free but have to buy a subscription for additional 'premium' features.

Twitch is a global live-streaming video service. Its content centres on streams of video games, including eSports competitions.

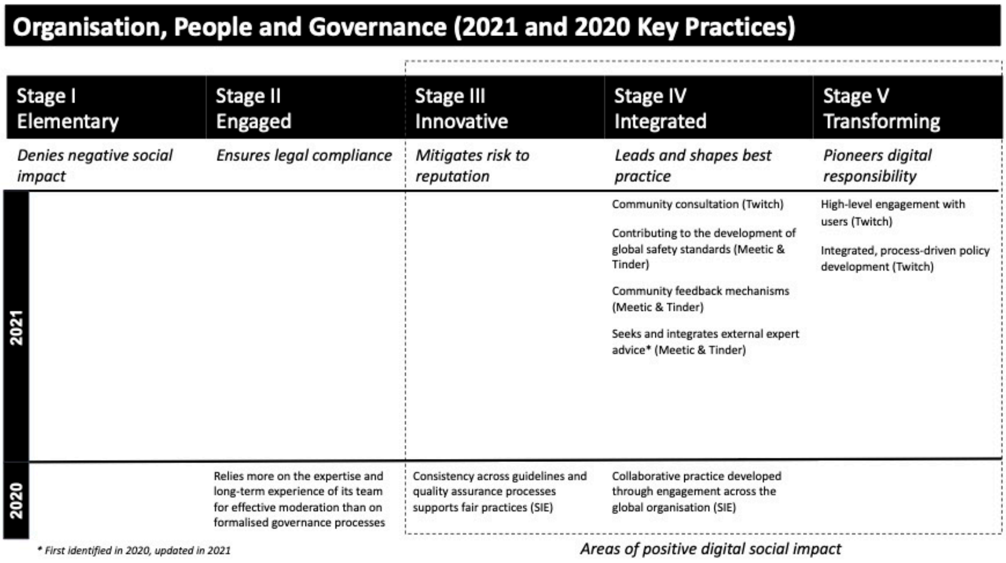
3.1 Organisation, people and governance

This section looks at an organisation's scope and purpose, the people concerned and its governance. We considered each organisation's mission, values and business model: why and how does it provide its service? We review how organisations connect and interact with stakeholders, including individual users and other organisations, and discuss how considerations of trust, safety and freedom are integrated into organisational culture and practice.

Overview

The extent to which organisations communicate with users through structured processes and feedback mechanisms can indicate their maturity with respect to digital responsibility. For instance, organisations that engage users when crafting policy and guidelines, and those that enable wider support communities, may build trust and confidence by giving a meaningful voice to users. Practices we observed include user surveys, focus groups, and forums. The most mature practices involve engaging users in high-level policy and product decision-making.

When addressing complex topics such as freedom of expression, inclusivity and mental health harms, third parties can help organisations to understand and prevent risks and harms. Less mature organisations do not formalise governance processes and instead implement policy without thorough testing or consulting with external experts.



Organisation, People and Governance (2021 and 2020 Key Practices)

	Stage I Elementary	Stage II Engaged	Stage III Innovative	Stage IV Integrated	Stage V Transforming
	Denies negative social impact	*Ensures legal compliance*	*Mitigates risk to reputation*	*Leads and shapes best practice*	*Pioneers digital responsibility*
2021				Community consultation (Twitch) Contributing to the development of global safety standards (Meetic & Tinder) Community feedback mechanisms (Meetic & Tinder) Seeks and integrates external expert advice* (Meetic & Tinder)	High-level engagement with users (Twitch) Integrated, process-driven policy development (Twitch)
2020		Relies more on the expertise and long-term experience of its team for effective moderation than on formalised governance processes	Consistency across guidelines and quality assurance processes supports fair practices (SIE)	Collaborative practice developed through engagement across the global organisation (SIE)	

** First identified in 2020, updated in 2021*

Areas of positive digital social impact

Adapted from Jonathan E. Shipp (2019) and Gluszek, Ewa. (2018) CSR Maturity Model – Theoretical Framework. Journal of Corporate Responsibility and Leadership. 4. 25. 10.12775/JCRL.2017.015.

Figure 3

Stage V (Transforming): Pioneers digital responsibility

3.1.1 High-level engagement with users *(newly identified 2021)*

Twitch includes several high-profile content creators in its Safety Advisory Council. The Council was established to inform product and policy decisions and highlight the potential impacts on marginalised people. By including both online safety experts and the service's content creators, who deeply understand trust and safety challenges on the service, Twitch synthesises academic input with practical experience and better informs the development of safe environments online. Moreover, engaging content creators at this high level formalises the relationship between the organisation and its users and empowers the user community. Twitch advances digital responsibility by integrating community input into its organisation structure, addressing disconnections between the service's governance policies and the practical realities of its products and policies for the user community.

3.1.2 Integrated, process-driven policy development *(newly identified 2021)*

Twitch incorporates the perspectives of multiple stakeholders across its teams and internal groups in policy development. This has led to an expansive policy set pertaining to the decisions behind the expected impact of new policies and products. Various interests across the organisation are brought together to review, socialise and role-play the impact of upcoming policies and products. Twitch makes effective use of its regularly updated blog to communicate policy updates and other developments which impact its users. Considering the substantial moderation challenges for live-streaming services, Twitch's approach to policy development is leading the way in carefully assessing the likely impact of its policies before their full release.

Stage IV (Integrated): Leads and shapes best practices

3.1.3 Community consultation *(newly identified 2021)*

Twitch is engaging with its user communities by investing in online spaces in which users feel free to voice their concerns and submit constructive feedback to influence decisions and updates on the service. In addition to these established forums, surveys are also shared with users to gather quantitative data about user experience, interviews and group sessions are conducted with users to garner insights about the community, and Twitch engages in 'listening' on social media to better understand their responses and broader sentiment around changes to policies and features.

3.1.4 Seeks and integrates external expert advice *(identified 2020, updated 2021)*

As subsidiaries of the same parent company, Meetic and Tinder benefit from its recently-formed group advisory council. The council includes experts and advocates involved in the study and prevention of sexual assault and harassment, sex trafficking and similar issues. In the US, Tinder has partnered with an anti-sexual violence organisation to inform their thinking around their reporting, moderation, and response policies and procedures. These improvements continue to be rolled out across the portfolio. In France, Meetic refers users to a loneliness prevention charity if they appear to pose a threat to themselves. This complements the strategic integration of external expertise through the group advisory council. Seeking external guidance on approaches to safety at the intersection of digital and real-world interaction contributes to protecting users from harm both on and off the service.

3.1.5 Contributing to the development of global safety standards *(newly identified 2021)*
The parent company of Meetic and Tinder aims to play an active role in defining the standards for all players in its industry and the technology sector in general. It works closely with legislators and regulators across the globe to contribute to the agreement of new safety-focused standards and laws, to help make both its own and other internet users safe. It seeks to lead and shape the legislative and policy landscape, rather than wait to follow in the wake of legislation and implement measures as required by law.

3.1.6 Collaborative practices developed through engagement across the global organisation *(identified in 2020)*
SIE engages across its global organisation to develop collaborative practice: effective processes and a culture of cross-functional sharing of insights supports the delivery of well-informed safety by design. Regional operations are coordinated through a global safety summit and a quality steering group. Shared data about moderation decisions is used to inform policy and product teams, ensuring consistent guidelines and decision-making, supporting fair practices and potentially enabling the translation of intelligence into technology.

3.1.7 Community feedback mechanisms *(newly identified 2021)*
Meetic and Tinder incorporate user feedback into their safety development processes: surveys are used to evaluate user experience, with a particular focus on safety. Consulting with the user community helps to validate service design and builds confidence and respect among users who feel their voices are being heard.

Stage III (Innovative): Mitigates risk to reputation

3.1.8 Consistency across guidelines and quality assurance processes supports fair practices *(identified in 2020)*
For SIE, the even application of community guidelines across a diverse moderation estate is crucial to produce a sense of fairness and transparency in fast-moving gameplay environments. Rules, drawn up by team leaders and informed by executive-level strategy, are set out in a clear and engaging moderator's handbook and a detailed education framework for its users. Users and moderators are accountable to one another in enacting an open and consistently enforced set of rules, and a system of quality assurance routinely monitors and calibrates every moderator's performance. Regular reviews of the guidelines with an emphasis on their clarity help to promote a responsible online culture.

Stage II (Engaged): Ensures Legal Compliance

3.1.9 Relies more on the expertise and long-term experience of its team for effective moderation than on formalised governance processes *(identified in 2020)*

One organisation relies more on the expertise and long-term experience of its team for effective moderation than on formalised governance processes or external consultation. Their human processes are likely effective but not always easy to document and review. While automated moderation has reduced the incidence of human moderators seeing extreme content, and the platform accepts some risk of over-blocking in return for a generally safer environment, it could expand its evaluative procedures to help fine-tune the moderation filters and gain greater insight into the full range of content being posted to the platform.

3.2 Content moderation

This section explores how harmful and illegal content, contact and conduct is discovered and acted upon. It looks at the organisation's content moderation policies, guidelines and procedures and uncovers ways in which decisions about content moderation are made and communicated to users. It reveals how organisations report breaches of their rules and local laws, some of the ways decisions can be challenged, and what happens when they are. It also looks at the resources applied to content moderation, including those to support the wellbeing of moderators on the front line.

Overview

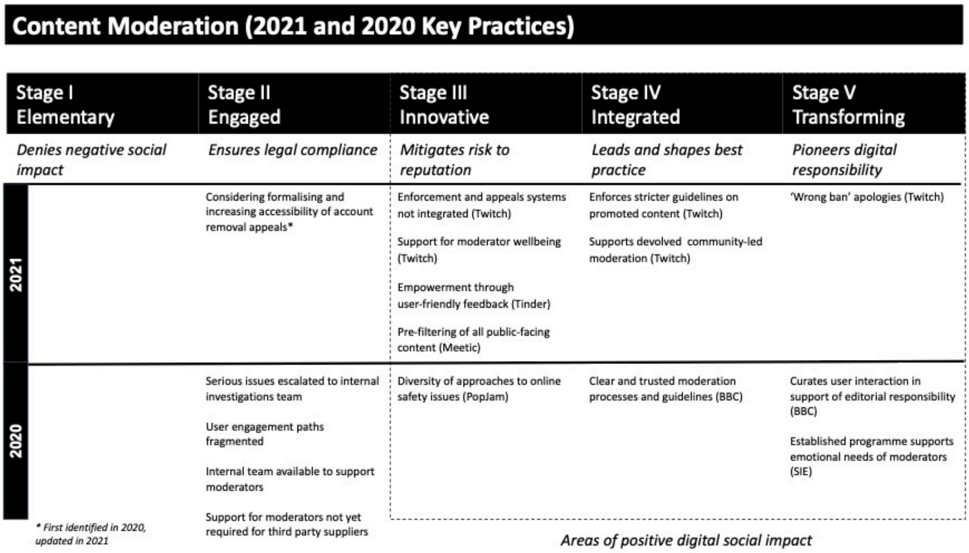
The mechanisms by which online services filter content, contact and conduct are generally unseen and unknown. This section reveals how organisations enable bottom-up moderation by empowering user reporting and facilitating community moderators, how organisations ensure the quality and welfare of their moderators, and the efforts taken to protect freedom of expression and build trust in the digital environment.

Although organisations are always able to control their content, the appropriate systems and processes vary with the nature of the service provided. In cases where organisations do not monitor behaviour, such as in private messages or off-platform, users can still be empowered and encouraged to report malicious content. In one case, where scale and ephemerality make it challenging to provide moderation, users are given tools with which to moderate their own micro-communities. As well as tackling a practical problem, this approach may help to achieve a sense of shared ownership and responsibility for the safety and integrity of online spaces.

There are a variety of approaches to ensuring the wellbeing of moderation staff and the quality of their work. Organisations that demonstrate leadership in this area provide front-line moderators with the necessary tools to face disturbing content daily: individual and group counselling sessions; ongoing monitoring and wellness programs; and provision of information and resources. This both alleviates harm to staff and improves the quality of moderation, by retaining valuable skills and experience. Pioneers of digital responsibility extend and standardise these practices to include third-party moderators. Less mature practices rely on moderators to come forward on their own initiative when suffering from exposure to disturbing content. In these cases, moderators may fear the professional consequences of voicing concerns when not encouraged to do so.

The overall processes and systems by which decisions are taken about content can be indicative of an organisation’s business culture, perhaps more so than the specific content policies applied for the assessment and removal of content. The way organisations communicate moderation decisions, apologising for incorrect decisions and build transparent appeals processes, illustrates the extent to which ethical considerations are embedded into an organisation’s operations.

Stage V (Transforming): Pioneers digital responsibility



Content Moderation (2021 and 2020 Key Practices)

	Stage I Elementary	Stage II Engaged	Stage III Innovative	Stage IV Integrated	Stage V Transforming
	Denies negative social impact	*Ensures legal compliance*	*Mitigates risk to reputation*	*Leads and shapes best practice*	*Pioneers digital responsibility*
2021		Considering formalising and increasing accessibility of account removal appeals*	Enforcement and appeals systems not integrated (Twitch) Support for moderator wellbeing (Twitch) Empowerment through user-friendly feedback (Tinder) Pre-filtering of all public-facing content (Meetic)	Enforces stricter guidelines on promoted content (Twitch) Supports devolved community-led moderation (Twitch)	'Wrong ban' apologies (Twitch)
2020		Serious issues escalated to internal investigations team User engagement paths fragmented Internal team available to support moderators Support for moderators not yet required for third party suppliers	Diversity of approaches to online safety issues (PopJam)	Clear and trusted moderation processes and guidelines (BBC)	Curates user interaction in support of editorial responsibility (BBC) Established programme supports emotional needs of moderators (SIE)

** First identified in 2020, updated in 2021*

Areas of positive digital social impact

Adapted from Jonathan E. Shipp (2019) and Gluszek, Ewa. (2018) CSR Maturity Model – Theoretical Framework. Journal of Corporate Responsibility and Leadership. 4. 25. 10.12775/JCRL.2017.015.

Figure 4

3.2.1 Established programme supporting emotional needs of moderators *(identified 2020)*

SIE's mandatory psychological support programme for moderators extends to all moderators, including those working for third-party suppliers. It includes monthly group, and quarterly individual, counselling sessions with a wellness provider. This ensures that all team members, including those who may find it difficult to seek help voluntarily, receive counselling support. Access to trained counsellors is available beyond the mandatory sessions, and moderators can seek help at any point. Counsellors can see the issues that moderators are escalating, and so may at times be more proactive in their support. This strong commitment to moderator welfare is likely to result in a higher performance, more caring and positive culture in the organization, as people's experience is retained and developed.

3.2.2 'Wrong ban' apologies *(newly identified 2021)*

Twitch has implemented an apology mechanism for users who are found, via the appeals process, to have been wrongfully banned. Communicating with users in this way promotes a shared sense of accountability. In pursuit of further transparency, users are also sent (pre-written) emails concerning the progress of their appeal, which will be supplanted in the coming year with a dashboard for appeals that will contain suspension-specific updates. Users who believe they have been unjustly banned will be more likely to pursue the appeals process if they consider errors to be openly acknowledged and corrected.

3.2.3 Curates user interaction in support of editorial responsibility *(identified 2020)*

The BBC proactively identifies areas of its website that would benefit from and support online user interaction. Once an area is open for user comments, the internal moderation management team invites and curates contributions. Opening and closing threads are key decisions that can promote sustained user engagement with new content continually. Moderation practices are designed to encourage a participatory culture, comments are treated with care and respect, and decisions err on the user's side, focusing first on users' intentions when reaching a judgement about the suitability of their posts. The BBC has promoted a culture of openness and respect that encourages public debate and upholds the value of audience input into key issues.

Stage IV (Integrated): Leads and shapes best practices

3.2.4 Supports devolved community-led moderation *(newly identified 2021)*

Twitch's devolution of limited moderation to user communities enables creators to facilitate online sub-cultures adapted to their streams. Creators can appoint trusted users to act as channel moderators, 'Mods'. Mods set the level at which automated moderation tools filter content and can blacklist specific terms. Since the vast majority of content on the service is public ('one-to-many'), this layered approach of internal and community enforcement must operate coherently. The organisation has sought to empower community moderators whilst keeping devolved moderators in line with the organisation's broader standards for content moderation.

3.2.5 Enforces stricter guidelines on promoted content *(newly identified 2021)*

Twitch's content which is promoted on the service's front page carousel is subject to stricter guidelines than other content, and all such content is hand-selected. Where promotion leads to higher volumes of viewership, greater scrutiny must be afforded to the content being promoted. This shows an awareness of good practice, and mitigates the reputational risk of promoting inappropriate content. This practice demonstrates that Twitch has a keen awareness of its role as a public forum for a broad variety of users and its associated responsibility to prevent malicious actors gaining a greater audience.

3.2.6 Clear and trusted moderation processes and guidelines *(identified in 2020)*

The BBC approaches content moderation as an extension of its strong reputation for high-quality editorial content, applying similarly high standards to online user-generated content. It curates user interactions in support of a wider digital media offer, in which commitments to high-quality user contribution play an important role. The BBC's notice and appeals process aims to promote a culture of openness and encourage public debate. Decisions should favour users, treating them as trustworthy contributors, an approach that is reflected in guidelines that focus first on users' intentions when reaching a judgement about the suitability of their posts. The BBC demonstrates a positive social impact by promoting a culture of openness and respect that encourages public debate and upholds the value of audience contribution.

Stage III (Innovative): Mitigates risk to reputation

3.2.7 Empowerment through user-friendly feedback *(newly identified 2021)*

In certain cases where users have been found to be in breach of the organisation's community guidelines for a relatively minor offence, Tinder has implemented a strike system whereby instead of immediately banning a user, they issue a strike or warning against them and offer opportunities for feedback. For instance, offering private information in a biography is not allowed under Tinder's Community Guidelines but the user may not realise this, so Tinder offers information about the reason for the strike and explains to the user where they have gone wrong. In offering users opportunities to become more informed about infractions of rules, Tinder is offering a more equitable pathway to improving the safety of users, delegating agency to users, and cultivating a sense of responsibility that allows users to align their behaviour with Tinder's expectations before more severe enforcement action is required. The company has found a significant degree of success in this strategy, having seen very low rates of recidivism and a significant reduction in bans. User empowerment has been at the heart of the recent language update in reporting flows. The articulation of possible violations in user-friendly language which maps "real-life" harms onto policy language is enhancing the role users play in improving the quality of the service by better capturing their negative experiences on- and off-platform.

3.2.8 Diversity of approaches to online safety issues *(identified 2020)*

PopJam's experienced team demonstrates strength in the diversity of approaches leveraged to promote and model positive content and behaviour. Stability, cohesion and high levels of mutual trust enable them to apply their experience to lead positive behaviour. Automated triage prioritises the most urgent cases for human moderation, freeing moderators to act as active role models for the community, as well as tackling problems. This strategy to promote and model positive content and online behaviour engenders a wider culture of creativity and respect.

3.2.9 Support for moderator wellbeing *(newly identified 2021)*

Twitch offers robust wellness support to its internal moderators and works to proactively reduce moderator exposure to disturbing content. There is in-house counselling, a 24/7 emotional support app, and access to emotional assistance programmes. Moderators are rotated through a queue of varying types and severities of reports, breaks are structured and flexible time is offered. Twitch's approach to vendor moderator wellness is similarly strong. Resources offered to third-party moderators are, however, configured slightly differently to support managed scaling with moderator resources and supply chain monitoring of moderator support. To be a leader in this area, Twitch could mandate elements of its support and fully extend its offering to external moderators, ensuring consistency in moderator support across the supply chain.

3.2.10 Pre-filtering of all public-facing content *(newly identified 2021)*

Meetic mitigates the risk of harm by automatically pre-filtering all publicly available photos and user descriptions before being published, so inappropriate profile pictures and biographies are identified and removed and, in serious cases, malicious actors are detected and removed.

3.2.11 Enforcement and appeals systems not integrated *(newly identified 2021)*

On Twitch's service, users cannot connect an appeal with a specific enforcement action and moderation staff must spend time checking across the two systems to validate the appeal. This disconnected approach has negative impacts for both users (who may struggle to appeal an enforcement action) and moderators (who are subjected to greater manual tasks creating longer response times). This has the potential to allow questionable – or simply incorrect – moderation decisions to simply go unchallenged. Twitch is aware of the issues posed by the friction between the two systems and aims to release an integrated enforcement and appeals system soon.

Stage II (Engaged): Ensures Legal Compliance

3.2.12 Considering formalising and increasing accessibility of account removal appeals *(identified 2020, updated 2021)*

On one organisation's service, there is no formal right of appeal except in the case of underage users and spam; users must email the customer services team to request a review. The organisation is looking at extending the formal right of appeal in a resource-efficient way. It has already expanded its automated remediation pathways from being only available to those users whose accounts have received an age-gate suspension, to including users who have been caught in the spam filter. These users are now able to complete a CAPTCHA to prove that they are human (and not a bot) and have their account reinstated.

3.2.13 User engagement pathways are fragmented *(identified 2020)*

One organisation's user engagement paths were fragmented and sometimes opaque. Clearer paths can make user appeals easier. The organisation is beginning to implement a new web-based appeals system which should enhance the process.

3.2.14 Internal team available to support moderators *(identified 2020)*

One organisation delivers their human moderation using a third-party supplier with two bases, in two different countries. Staff are trained and managed by an internal team, who are available to support moderators. Mental health support is a provision that is embedded in the management process, whereby supervisors are expected to operate an open-door policy to allow moderators to raise issues or concerns they have with them.

3.2.15. Serious issues escalated to internal investigations teams *(identified 2020)*

In one case, reports of illegal or at-risk content are normally escalated internally. However, for content on an external service, responsibility for measures of protection and intervention lies first with that service, so issues are usually reported to the partner rather than escalated internally. This may not always match users' expectations.

3.2.16. Support for moderators not yet required for third party suppliers *(identified 2020)*

Moderation can be a mentally and emotionally challenging task. In one case, although support is available, counselling for moderation teams is not yet required as part of procurement agreements with third party suppliers. This is, however, something that the organisation is working towards.

3.3 Automation

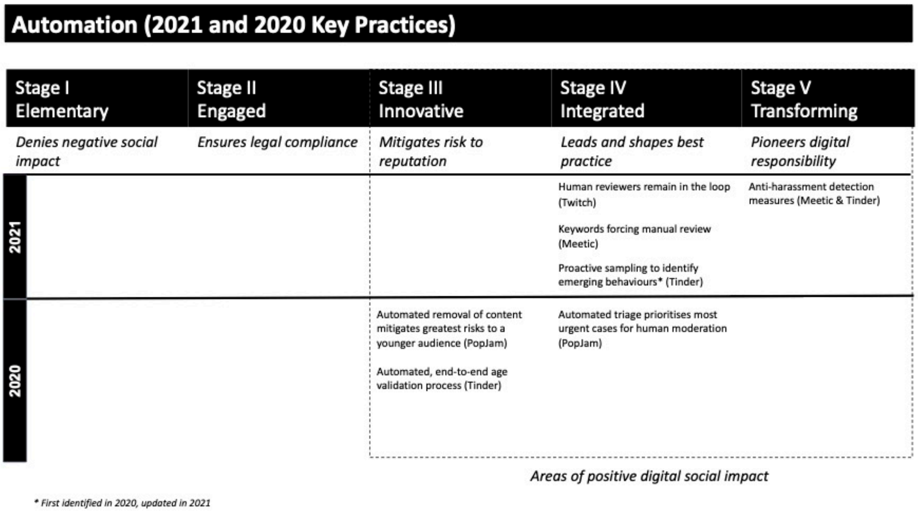
This section looks at how intelligent and automated systems are used to shape service design, promote and moderate online content and keep users safe. It reviews how behavioural analytics is used to shape user experiences, explores how and why automation is deployed in content moderation, and the extent of human oversight involved.

Overview

The key challenge for companies using automated processes to create safe online environments is to ensure that the right checks and balances are in place. Automated decision-making technologies and processes need to be regularly updated, flexible and subject to human oversight: if not, they replicate existing decision-making, whether it is right or wrong. The choice of training data, to build such systems, is an important consideration when seeking to avoid replicating and exacerbating existing bias.

Large-scale and persistent threats require automated responses. Tools based on Artificial intelligence (AI) and Machine Learning (ML) technologies can efficiently triage and prioritise the most urgent content moderation cases so that child sexual abuse material is removed before any exposure to users. To give a sense of the scale, one organisation's user base of 140 million produced 19 billion messages in one quarter alone and 18 billion hours of live video content over a year. AI and ML technologies can also uncover malicious content and new types of threats. One organisation uses automation to detect behaviours that may indicate potential harassment, nudging users to report abuse if appropriate. It is standard practice to fully automate removals when detecting networks of bots, spam and persistent fraud.

Relying entirely on automation to moderate dangerous content may, however, create risks to freedom of expression. Benign actors may be incorrectly punished in contexts not easily interpreted by tools based on AI and ML technologies. The use of these technologies to quickly identify and remove known child abuse images may be easy to justify, but the censoring effect of incorrect removals of contentious opinions or political speech is significantly harder. In recognising the limitations of automation, one organisation insists on human review for all removal decisions; another uses keyword triggers to manually review context-sensitive content especially prone to machine error.



Automation (2021 and 2020 Key Practices)

	Stage I Elementary	Stage II Engaged	Stage III Innovative	Stage IV Integrated	Stage V Transforming
	Denies negative social impact	*Ensures legal compliance*	*Mitigates risk to reputation*	*Leads and shapes best practice*	*Pioneers digital responsibility*
2021				Human reviewers remain in the loop (Twitch) Keywords forcing manual review (Meetic) Proactive sampling to identify emerging behaviours* (Tinder)	Anti-harassment detection measures (Meetic & Tinder)
2020			Automated removal of content mitigates greatest risks to a younger audience (PopJam) Automated, end-to-end age validation process (Tinder)	Automated triage prioritises most urgent cases for human moderation (PopJam)	

Areas of positive digital social impact

** First identified in 2020, updated in 2021*

Figure 5

Stage V (Transforming): Pioneers digital responsibility

3.3.1 Anti-harassment detection measures *(newly identified 2021)*

Meetic and Tinder have implemented anti-harassment measures, deploying nudges on both ends of user communications. The two organisations approach the issue of harassment differently, partially on account of their respective legal environments; nevertheless, both Meetic and Tinder show leadership by pioneering innovative solutions to their most pertinent challenges.

Meetic seeks to shape user behaviour, helping them to feel safe and empowered by encouraging reporting. It has been working to identify patterns of user behaviour that might suggest misconduct. In particular, Meetic has identified that when a certain number of messages are sent without any response from the recipient, it can indicate potential harassment. When this threshold is reached, an advice message is shown to both the sender and recipient of the messages. This is intended to encourage the recipient to report harassment so that moderators can review and take enforcement action where appropriate.

Tinder uses ML-based tools to scan private messages and detect and flag anything potentially harmful or inappropriate. In Tinder's specific context, individual preferences and other factors can affect how a comment is intended or received in a way that ML cannot always detect. The tool was therefore designed to first ask the recipient if they perceive a flagged message to be harassment, to direct the user to report if so. This measure also incorporates user feedback in the moderation process as content reported as harassing is, after being confirmed by a moderator, used to further refine the tools. This feature has generated significant insight for Tinder about what may constitute offensive or harmful content and has assisted the development of new prompts; for instance, Tinder has recently rolled out a feature whereby senders are prompted to consider whether their message might be perceived as harassing before sending.

Stage IV (Integrated): Leads and shapes best practices

3.3.2 Human reviewers remain in the loop *(newly identified 2021)*
Twitch deploys AI for flagging content, not for decision-making regarding whether content should be removed. Human review is an integral part of the moderation process across the service. Suspensions are not issued without human review and there are no plans to change this. ML tools are primarily used to flag items for human review, except in cases of banned expressions (specific usernames, for example), or for issues involving waves of bots. The username tool is preventative, operated as a gate at the point of sign-up, is re-trained daily, and is constantly validated with human review. The AI deployed in Twitch's automated moderation tools is provided by a third-party vendor. Twitch staff meet weekly with them to develop the system, informed by community feedback. This maximises the utility of automated moderation tools whilst using human oversight to ensure they are tackling the issues encountered by users and identifying emerging problematic trends.

3.3.3 Keywords forcing manual review *(newly identified 2021)*
Meetic's rules for automated detection systems mean that specific words or phrases known to relate to sensitive topics or issues will force a review by a human moderator. This enables context to be considered in cases that automated systems may not be able to handle well. Given that it is a closed environment enabling peer-to-peer contact rather than an open, one-to-many type service, Meetic does not consider freedom of speech to be a major issue on its service. Nonetheless, consideration of context will help to ensure that individual moderation decisions are fair and that training sets for automated decisions are balanced.

3.3.4 Proactive sampling to identify emerging behaviours *(identified 2020, updated 2021)*
Tinder deploys ML to proactively check all public-facing content on its service with some limited review of private messages, moderation of which is generally only reactive, triggered by user reports. Possible violations are flagged for the first round of human review and potential removal. Notably, Tinder has tuned its automated systems to overestimate the amount of content in breach of guidelines, maximising threat detection by flagging around 25% of image content as a potential violation. The system also intentionally flags a random sample of content for review by human moderators looking to identify new and emerging patterns of abuse. This casts a wider safety net around the moderation process, ensuring that the majority of content violations are removed before being viewed by users.

3.3.5 Automated triage for most urgent cases *(identified 2020)*
In partnership with a third party technology provider, PopJam supported the design of an automated system that prioritises the most urgent cases for human moderation. After the initial removal of high-risk content by the automated moderation tools, the supplier's triage system sorts the remaining cases into a hierarchy of moderation queues according to how risky each item of content is. The system

predicts the likelihood of problematic material by type of content, with the most pressing and urgent placed at the top. This process helps to reduce the risk of a very urgent piece of content getting stuck behind a long queue of more minor issues. The system is resilient and serves to increase the efficiency of the moderators. A team leader liaises with the third-party supplier and is involved daily in the managing of the triage process. Low priority queues wait until the site closes at night and the moderators can deal with any backlog.

Stage III (Innovative): Mitigates risk to reputation

3.3.6 Automated, end-to-end age verification process *(identified 2020)*
Tinder is an exclusively 18+ service, and age-gating measures are built into the registration process. Any user entering an underage date of birth sees their credentials blocked until they turn 18 according to the date of birth entered. ML-based tools are used to detect underage users through their photographs, biographies and private messages. Once a user is suspected of being underage, their account is suspended and can only be reinstated once an age verification process has been completed.

3.3.7 Automated removal of content mitigates greatest risks to a younger audience *(identified 2020)*
PopJam uses a system of automated removal, whereby potentially egregious content is immediately suppressed. For instance, child abuse images and references to self-harm are capable of severely impacting the organisation's young user community. Considering the imitative potential of such behaviour, immediate and outright removal of any threat is prioritised. In doing so, PopJam could signpost support to users and learn from negative behaviour patterns. On balance, however, the suppression of the imitative negative content and the promotion of space for positive engagement is likely a net gain for its audience.

3.4 Safety

This section looks at the protection of users' health and well-being. It considers how organisations take responsibility for user welfare, how services are designed with safety in mind, and how incentives are used to foster safe online environments. We also consider children's experiences of online environments and how their interests are considered in service design and development processes.

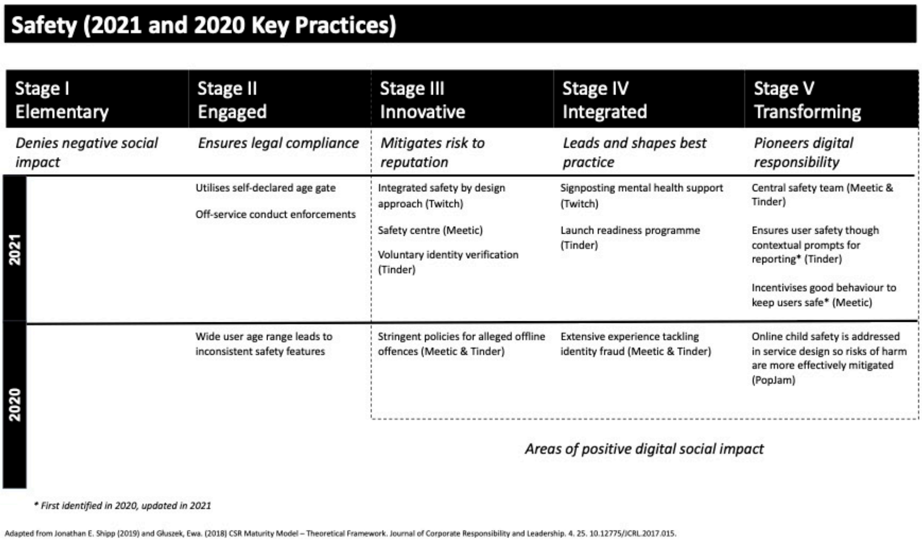
Overview

All the participating organisations considered there to be some inevitable trade-offs between safety, privacy and freedom of expression. In some instances and jurisdictions, a policy is constrained by legislation, for instance as regards encryption and access to private messages. In others, the nature of the service dictates what policy and strategy may be appropriate: a dating service will be designed to facilitate real-life meetings between users, whereas a service designed for children may actively try to stop this from happening.

Novel practices nudge users to report potential violations, empowering them to judge the context of situations rather than relying on automated tools that create risks to freedom of expression when content is mistakenly removed. One organisation encourages users to verify their identity but makes the process voluntary to respect privacy.

For children's services, safety is the principal consideration. In some cases, private messages are scanned despite the privacy implications. In others, communication functions are limited, justified on safety grounds despite the limitation on freedom of expression. One service moderates content for a broad range of ages but is yet to devise the layered approach which might be necessary to do this successfully. For any service provider, it is important to seek ways to harmonise safety across different user types.

Considering potential harms and possible trade-offs at the earliest stage in online service design is an important approach. Safety by design is increasingly expected and deployed within organisational practices and may by now be considered standard practice.



Safety (2021 and 2020 Key Practices)

	Stage I Elementary	Stage II Engaged	Stage III Innovative	Stage IV Integrated	Stage V Transforming
	Denies negative social impact	*Ensures legal compliance*	*Mitigates risk to reputation*	*Leads and shapes best practice*	*Pioneers digital responsibility*
2021		Utilises self-declared age gate Off-service conduct enforcements	Integrated safety by design approach (Twitch) Safety centre (Meetic) Voluntary identity verification (Tinder)	Signposting mental health support (Twitch) Launch readiness programme (Tinder)	Central safety team (Meetic & Tinder) Ensures user safety though contextual prompts for reporting* (Tinder) Incentivises good behaviour to keep users safe* (Meetic)
2020		Wide user age range leads to inconsistent safety features	Stringent policies for alleged offline offences (Meetic & Tinder)	Extensive experience tackling identity fraud (Meetic & Tinder)	Online child safety is addressed in service design so risks of harm are more effectively mitigated (PopJam)

Areas of positive digital social impact

** First identified in 2020, updated in 2021*

Adapted from Jonathan E. Shipp (2019) and Gluszek, Ewa. (2018) CSR Maturity Model – Theoretical Framework. Journal of Corporate Responsibility and Leadership. 4. 25. 10.12775/JCRL.2017.015.

Figure 6

Stage V (Transforming): Pioneers digital responsibility

3.4.1 Ensures user safety through contextual prompts for reporting *(identified 2020, updated 2021)*
Tinder has considered the interpersonal interaction allowed on the service, limiting higher-risk features including, until recently, initiating a private video call. The group-level advisory council was consulted about how to best create a safe environment in the recently launched one-to-one video chat function, which responded to user interest and the limits to physical meetings resulting from the Covid-19 pandemic. Tinder designed this new feature with safety in mind built-in, end-to-end: users must have already connected with one another and agreed to a call before the feature is unlocked; they must recommit to following the guidelines; and after the call, they are asked whether they would use the feature again, and are immediately offered the chance to report an unsafe interaction. This 'reporting by design' helps assure users of Tinder's commitment to their safety.

3.4.2 Online child safety is addressed in service design so risks of harm are more effectively mitigated *(identified 2020)*
PopJam implemented its transformational practice at the service design stage. Capitalising on a long experience of online communities for 7-12-year-olds, PopJam designed out known opportunities for poor behaviour or taking advantage of user vulnerability. Private messaging, which is perhaps less valued by younger users, was not enabled, thereby preempting the risks of grooming or obscenity. PopJam demonstrates positive social impact by effectively mitigating child safety risks within its service design.

3.4.3 Central safety team *(newly identified 2021)*

Over the last two years, Meetic and Tinder's parent company has created and built up a centralised safety team. This functions as a central knowledge hub to oversee and enhance the individual safety functions of brands across the group's portfolio as well as identify and transfer best practices. The team is still growing and aims to incorporate a range of competencies from across the brands to centralise and coordinate an overarching strategy. It will include staff from both product and operations functions, as well as staff whose role is to deal more directly with (1) escalations and moderation, (2) partnerships with nonprofits and law enforcement and (3) international safety nuance. This initiative coordinates and aligns resources and approaches to safety while implementing an overarching technology governance structure.

3.4.4 Incentivises good behaviour to keep users safe *(identified 2020, updated 2021)*

Previously, Meetic incentivised and rewarded good behaviour using a trust badge. The badge was shown on the profiles of men who had followed video tutorials, had committed to a 'gentleman's charter' standard of good behaviour, and had a high level of detail in their profile. The badge was designed to reinforce good behaviour and highlight men who have made a visible commitment to courteous behaviour. Meetic was encouraged to see that men who have the badge appear to have more positive interactions with women. In 2021, noting some positive results of the badge, Meetic is developing alternative approaches to shaping positive behaviours for all users.

Stage IV (Integrated): Leads and shapes best practices

3.4.5 Launch readiness programme *(newly identified 2021)*

Tinder has formalised an internal set of processes whereby cross-functional stakeholders remain engaged with, and informed about, new product developments and updates. The programme is run as part of safety operations and allows safety specialists to inform the thinking of product managers as they develop features and updates. Biweekly meetings are held in which members of various teams can raise potential issues before they are designed into features. This also informs the creation of public-facing feedback tools and resources for users with regard to changes to the product, and internal briefings for moderators and support agents to update their practices accordingly. The process alleviates the burden on individual product managers to inform each stakeholder they think may need to be informed while maintaining open dialogue and enabling alignment across teams.

3.4.6 Extensive experience tackling identity fraud *(identified in 2020)*

Romance fraud causes significant financial and emotional distress and is one of the priority issues that the parent company of Meetic and Tinder tackles every day, to provide safe and enjoyable experiences for users. Sometimes called catfishing, fake profiles are created to ingratiate with and earn the trust of a user, before asking for or extorting money. Such fraud can lead to significant emotional distress and financial loss. Device identities, geographic location and user profiles are key data points for Tinder's fraud detection technologies, whose algorithms detect hundreds of patterns previously seen in scam and fraud accounts.

3.4.7 Signposting mental health support *(newly identified 2021)*

Twitch provides mental health support through signposting of relevant resources. This was done by partnering with a mental health service provider to develop the messaging they provide to users with mental health needs, as surfaced to the organisation in reports. Users may also text the name of the organisation to the mental health service provider to be connected with a counsellor immediately.

Stage III (Innovative): Mitigates risk to reputation

3.4.8 Voluntary identity verification *(newly identified 2021)*

In August 2021, Tinder announced the rolling out of voluntary identity verification. As it determines how the feature will roll out, Tinder will take into consideration expert recommendations, input from its members, the documents most appropriate in each country, and local laws and regulations. The product aims to balance user safety with a privacy-friendly approach to identity verification. The voluntary character of this practice comes with the limitation of transposing responsibility to the user regarding their engagement with “unverified” users.

3.4.9 Stringent policies for alleged offline offences *(identified 2020)*

Meetic and Tinder seek to create meaningful, safe, in-person encounters between users. If, for example, following a user meeting, a complaint is received about someone’s behaviour, the organisations act in support of the complainant and ban the reported user from the service. This stringent approach does allow the possibility of malicious reporting but Meetic and Tinder see a higher risk in allowing potentially bad actors to continue using their service.

3.4.10 Integrated safety by design approach *(newly identified 2021)*

Twitch has integrated safety by design. Its preventive, risk-based approach integrates safety considerations into the organisation’s product and policy development workflows. This involves training project managers in safety, requiring all specifications to be reviewed to identify possible safety risks, setting positive expectations for user safety, and aligning safety priorities across functional teams. Embedding multi-stakeholder review into the product development process demonstrates Twitch’s commitment to ensuring a sense of shared responsibility for the safety and wellbeing of users across the organisation. The provision of training and resources for those employees who do not specifically work in trust and safety or wellbeing is a proactive measure in aligning the goals of various teams on the user-centric approach that Twitch has taken.

3.4.11 Safety centre *(newly identified 2021)*

Meetic is developing a user safety centre based on the successful roll-out of a similar tool by Tinder, which has safety centres in almost every market. Meetic’s centre will improve user experience, tailoring it for each market, and making existing safety tools and advice more consistent, including signposting relevant advice from third parties. It is expected to be available to users in Q1, 2022.

Stage II (Engaged): Ensures Legal Compliance

3.4.12 Wide user age range leads to inconsistent safety features *(identified 2020)*

One organisation has struggled to manage user expectations due to its wide user age range. When the organisation is assessing potential new safety features, the broad range of users prevents a consistent, appropriate approach from emerging. While the publisher-moderated environments come with their moderation requirements and relevant age-appropriate rules, the organisation's responsibility is to create a welcoming environment for users of all ages. User content is moderated to match the age range determined for users on the organisation's interactive spaces, which is available for its youngest users. This may not feel at all congenial for other, older users in the same space, who are used to having more freedom to express themselves together. Moderators may sometimes feel pulled both ways by the diverse range of ages in the user community.

3.4.13 Utilises self-declared age gate *(newly identified 2021)*

The 13+ age-gate currently in operation on one organisation's service involves a self-declaration of age. If a user inputs a date of birth indicating that they are under the age of 13, a cookie will be dropped to temporarily restrict them from creating an account with another date of birth. 21+ age gates are also in operation on channels that feature promotions or sponsorships by alcohol brands. This alcohol-specific age-gate functions similarly in that it simply requires a self-declaration of age.

3.4.14 Off-service conduct enforcements *(newly identified 2021)*

One organisation recently announced a new policy whereby users can be suspended as a result of off-service behaviour. Behaviours that would result in this sort of enforcement action are largely limited to the most severe, such as participation in terrorist activities, membership in a hate group, sexual assault and anything related to child sexual abuse. If replicated or highlighted on the service, such behaviour could negatively impact service users and risk reputational damage to the organisation. At the point of evaluation, this practice was at an early stage in its maturity. It should be reviewed as more evidence becomes available.

4. Analysis

We are living through a global information crisis, characterised by an asymmetry of information between the public, policymakers and business organisations. Digitalisation brings significant trust and safety issues, to which adequate policy and regulatory responses are only just starting to be established. Independent and evidence-based identification and analysis of organisational practices is a vital first step in the development of the standards and benchmarks that will be needed to give digitalisation a new direction.

Some of the observed practices shape the online environment directly, by building or limiting user agency, raising expectations and setting standards for individual behaviour. Other practices are more internal, building and integrating an organisation's capacity for ethical business practice and leadership across various business operations. These internal processes are at the heart of how organisations act and operate as part of society. Our analysis should enable a greater understanding of how digital environments are being shaped, and the ethics of the organisations involved.

Shaping positive online experiences
Clear and open communication with users supports a participatory ethos and reinforces structured user engagement in setting and enforcing community standards. Collaborating with users builds a sense of shared responsibility for the community's safety and wellbeing.

By including users in stakeholder advisory groups, organisations can establish user agency and systematically incorporate user experiences and expertise into service design, further engendering positive user behaviour and building trust between user and service. At the same time, these efforts are undermined by fragmented or undefined reporting and appeals systems: any lack of user agency can manifest in harmful behaviours in the wider digital environment.

Not everything can be left to users to decide, however: organisations make a difference by enforcing stricter guidelines for higher-profile content and acting decisively to disrupt the proliferation of harmful and illegal content. Organisations must iterate their approaches to shaping good behaviour and promoting cultures of safety and respect. This can be particularly challenging where they aim to accommodate users with different devices and internet access configurations.

Monitoring tools can encourage reports of harmful messages, and prompt senders to think again, improving user safety and wellbeing. Leading organisations recognise the importance of human oversight of automated systems, and use appropriate training data so that bias is not replicated and amplified. Visible and effective safety processes build trust and raise the expectations of users across the online environment. Compliance-driven, tactical deployment of safety technologies without ethical oversight, however, threatens freedom of expression online.

Building ethical business practice

When guided by external expertise, and a commitment to protecting users from emerging harms, organisations differentiate themselves in the market. Cross-functional collaboration can help teams share and learn, and build cultures of digital responsibility. This helps organisations anticipate ethical considerations at the product design stage and so avoid difficult and expensive retrofitting. Well-aligned organisations are able to offer consistent user support, building trust and confidence in an organisation's service as well as internet services more broadly.

The welfare of front-line staff, and the appropriate use of automation to support them, is a key ethical challenge. Leading organisations have raised the bar in respect of these issues, demonstrating a commitment to produce and provide extensive support resources. Child safety is a majorl concern for organisations operating online. In this domain, automated age verification systems play a vital role in ensuring the safety of the most vulnerable users.

Considering the risks of leaving humans out of the loop, leading organisations seek an appropriate synergy of human and automated systems, wherein humans retrain automated systems to improve outcomes for users. Identity, age assurance and anti-fraud processes help protect users from a range of unwanted, inappropriate, and harmful content, contact and conduct. To be effective, however, these require investment, commitment, and ongoing improvement.

The organisations involved in this report are part of a global ecosystem that includes wide-ranging and interdependent stakeholders. The orientation of this analysis towards the culture and values that drive decision-making contributes to the understanding of corporate purpose and ethical business practice. It empowers organisations to share knowledge and demonstrate ethical leadership, informs investors to better assess investment risks, enables civil society actors to better understand how corporate bodies implement policy, supports policymakers in addressing risks to citizens and consumers, and helps campaigners and researchers to keep asking the right questions.

5. Conclusions

Our work envisages Corporate Digital Responsibility in a holistic way that brings together corporate mission, organisational structures and business practices driven by societal purpose. We argue that organisational maturity is reflected in an organisation's strategic priorities, a systems and processes approach and the alignment of corporate mission and values with the real-world impact of services and products. Our analysis of organisational practices has identified three key themes:

1. User agency: engaging users in the strategic shaping of policy
Leading organisations strategically and systematically incorporate the voices of users when designing and developing their services. Positive online experiences are shaped by attending consistently and fairly to the needs of users, alongside the voices of experts and policymakers.

2. Proactive oversight: taking responsibility for social impact
Organisations can change the game by taking responsibility for the impact of their services, including at the highest levels in the organisation. They do this both in the design of their services and operations and through their participation in the development of safety standards. This oversight is especially important when deploying automation.

3. Innovation: new technologies can support both safety and freedom
Innovations in identity verification, age assurance and monitoring can provide solutions that protect the safety of users, whilst mitigating the effect on the freedom of others. Considerable commitment, investment and iteration is necessary to find the right solutions and to build trust with users and wider stakeholder groups. The right solutions will vary between different organisations and services.

Next steps

The Internet Commission faces a challenging task in a complex ecosystem, characterised by power and knowledge asymmetries between internet companies, governments and civil society. However, we have the opportunity to help organisations around the world to learn from experts and each other and to demonstrate their commitment to digital responsibility.

We work hard towards ensuring our independence and inspiring confidence in the integrity of our process. We believe this report is significant: it offers a unique inside view about processes and decisions that determine the quality of the digital environment, which is now the foundation of cultural, political and economic life.

Our reporting cycles aim to build our capacity to deliver an effective accountability mechanism that contributes to the protection of fundamental freedoms. A third cycle will see the Internet Commission scaling its work, as well as further iterating its framework and methodology. We aim to build up the Internet Commission's bank of case studies and practices as a foundation for the identification, evaluation, and sharing of best practices.

We wish for this work to spark constructive conversations and look forward to engaging in them. Your feedback will help us evolve and improve.

Afterword

Jeremy West
OECD

The Internet Commission's Accountability Report 2.0 does what every instalment in a promising young series of evidence-based reports should do: evolve, improve, and add to our knowledge about an important issue. This edition marks another advance in understanding the state of companies' digital accountability, the progress made, and the directions in which future efforts should go.

The underlying Evaluation Framework provides standardisation and structure for the Report's evidence-gathering approach, enabling the cross-cutting Overviews of the four main Findings. The Overviews analyse patterns, similarities and differences along various dimensions of accountability practices and criteria.

Furthermore, by characterising the state of each companies' practices in the context of a maturity model that runs from "Elementary" to "Transformational", the report builds in a soft incentive for the participating companies to make improvements over time, while recognizing the progress they have already made. This is no shaming exercise. Instead, it is a demonstration of what multi-stakeholder cooperation, trust, and goodwill can achieve.

This approach resonates with the Organisation for Economic Co-operation and Development. The OECD is an international organisation that works to build better policies for better lives. Our goal is to shape policies that foster prosperity, equality, opportunity and well-being for all. Drawing on more than 60 years of experience and insights, the OECD is a trusted forum where governments, along with business, civil society and other stakeholder groups, come together to collaborate, establish evidence-based international standards and find solutions to a range of social, economic and environmental challenges.

Internet governance, including the challenges presented by illegal and "awful but lawful" content online, is among the policy areas in which the OECD works to improve the evidence base and provide analysis to support policymakers. Our new Voluntary Transparency Reporting Framework (VTRF), which was designed in part to increase accountability for Internet safety and human rights, aligns with many of the same principles and values that underlie this report. Focusing initially on transparency around terrorist and violent extremist content online, the VTRF is an international, standardised framework for transparency reporting that any content-sharing service can use and that all OECD member countries support.

We are now transforming the VTRF into a pilot web portal, oecd-vtrf-pilot.org, which we expect to launch in April 2022. That will mark the start of a period in which we will gather feedback from the companies that voluntarily complete the reporting questionnaire, and from the users who view the reported information. We encourage all stakeholders to help us make improvements by trying out VTRF 1.0 and telling us about their experience with it.

OECD
BETTER POLICIES FOR BETTER LIVES

Jeremy West is a Senior Policy Analyst in the Division for Digital Economy Policy at the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the Member countries of the OECD.

Afterword

Beeban Kidron

5Rights

If there is one thing we have learned in 2021 following revelations from former tech company employees turned whistle-blowers, it is that a company's ways of working impact on the lived experience of those who use their services. This is a truth self-evident for those of us campaigning for systemic changes across the tech sector, to ensure digital services are designed for safety and create a culture of transparency and accountability, and it is a view that sits behind the inquiries and conclusions of this report. Many of the specific indicators identified in this report, including clear roles and responsibilities, engaging meaningfully with user groups, risk assessments, mitigation strategies, pathways for escalation and keeping humans in the loop – to mention just a few – are consistent with other industries and sectors. These should not be considered innovative, but simply the price of doing business.

I therefore welcome the Internet Commission's latest accountability report, and in doing so would urge all companies to take on board each of the obvious examples of responsible design that it explores. Whilst I, in my role as a parliamentarian, will continue to advocate for smart regulation and rules of the road that reflect children's rights and needs, it is abundantly clear that companies could, as a matter of course, do the right thing now. As this report carefully argues, doing the right thing is the only long term, sustainable position for individual companies and the sector.

In September 2021, the Age Appropriate Design Code[11] came into force in the UK and with it a whole raft of significant safety upgrades to major services. These included the introduction of default safe search mode for under 18s, disabling features that allow adult strangers to directly message children, turning off auto-play or notifications for particular ages or at certain times of the day, and stronger enforcement of age restrictions through the use of age assurance technologies. Many people were astounded that these changes had not been made sooner or were not already the norm, and somewhat bewildered that regulation was required to impose this obvious level of responsible design. They are of course quite right.

Doing the right thing is not simply signposting or picking those elements of design or governance that are most convenient, but rather a commitment through the life cycle of a service or product to anticipate the needs and wellbeing of users. In our work with the Institute of Electrical and Electronics Engineers (IEEE), a significant group of engineers, policy makers, enforcement, academic and business people came together to create IEEE standard 2089-21[14], a voluntary standard that sets out the processes and circumstances necessary to design age appropriate services. It has much in common with this report, which in many ways serves as a case study for its approach. I urge all who read this to look at standard 2089-21 and see how it overlaps with and enhances the valuable work of the Internet Commission.

5RIGHTS FOUNDATION

Baroness Beeban Kidron is a crossbench peer in the House of Lords. She introduced the Age Appropriate Design Code into law, and is the Chair of 5Rights Foundation.

5Rights Foundation develops new policy, creates innovative projects and challenges received narratives to ensure governments, regulators, the tech sector and society understand, recognise and prioritise children's needs and rights in the digital world. Our work is pragmatic and implementable, allowing us to work with governments, intergovernmental institutions, professional associations, academics, and young people across the globe to build the digital world that young people deserve.

[11] Age appropriate design: a code of practice for online services – ICO

[12] IEEE 2089-2021 - IEEE Standard for an Age Appropriate Digital Services Framework Based on the 5Rights Principles for Children

Afterword

Lourdes Montenegro

World Benchmarking Alliance

Digital technologies are among the most influential to achieving the SDGs. From finance apps that provide unbanked access to financial resources for those in poverty, to internet services that help people connect and cooperate globally, digital technologies are recognized as cross-cutting enablers to accelerate the SDGs. However, they also have a potential to harm sustainable development and must be developed and deployed in ways that respect human rights and leave no one behind. For this to happen, corporate benchmarking must play a greater role. Effective benchmarking can both highlight companies that are leading the way, triggering a race to the top, and hold underachieving companies accountable.

In 2021, the WBA released its second Digital Inclusion Benchmark , to assess how companies are helping to advance an inclusive and trustworthy digital transformation. It is clear from the findings that the digital sector, especially digital platforms, have a long way to go. Only 27 out of 150 companies assessed achieved overall passing marks averaged across four areas: enhancing universal access, improving all levels of digital skills, fostering safe and trustworthy use, and innovating openly and ethically.

The Internet Commission therefore plays a vital role in helping drive a trustworthy digital transformation. By understanding how internet companies approach the impact of their business, they can encourage others to improve and hold accountable those that don’t. This report also uncovers practices and processes that themselves affect the SDGs. Effective content moderation systems, for instance, directly relate to Goal 3 – good health and wellbeing – and can both support healthy lifestyles and relationships and mitigate the threats to wellbeing such as online harassment. Practices that provide protections for front-line content moderators, too, align with Goal 8 – promote sustained, full and productive employment and decent work for all.

The Internet Commission’s work creates a better understanding of how internal corporate culture and organization facilitates or hinders digital responsibility. We are pleased to have the Internet Commission as a WBA Ally and we welcome others interested to collaborate with us and join our movement to drive corporate accountability for sustainable development.

Lourdes leads on digital sector transformation at the World Benchmarking Alliance (WBA), including strategic oversight of the Digital Inclusion Benchmark covering 200 of the most influential tech companies in the world.

The WBA is a global initiative providing an accountability mechanism for private sector contributions to achieving the UN's Sustainable Development Goals (SDGs). Echoing the true spirit of SDG17 – Partnerships for the Goals – we work with over 270 Allies worldwide to leverage the power of benchmarks and cross-sector partnerships to drive systemic progress on the SDGs.

Appendix 1: About the Internet Commission

As an independent, trusted broker within the new regulatory system, the Internet Commission aims to ask the right questions, provide reliable evidence and help organisations to navigate different national and international requirements. It offers independent health checks, knowledge sharing and review services to organisations that lead in digital responsibility, and authoritative insight to regulators and other stakeholders.

We are grateful to our partners for their continued collaboration and support.

- Arm
- Bates Wells
- Carnegie UK Trust
- Global Enabling Sustainability Initiative (GeSI)
- London School of Economics and Political Science
- Match Group
- Pearson
- Sony Interactive Entertainment Europe
- Twitch
- United Nations Secretariat
- Wayra
- WePROTECT Global Alliance

Appendix 2:
Evaluation Framework for Digital Responsibility

This Evaluation Framework looks at how organisational cultures, systems and processes align to support corporate digital responsibility. It has a particular focus on internet safety, freedom of speech and the ways in which decisions are made in relation to content, contact and conduct online.

It is designed to enable information to be shared with the Internet Commission in order to facilitate the creation of (a) confidential individual case studies (b) confidential knowledge sharing and (c) a published accountability report. Information shared with the Internet Commission will remain strictly confidential unless otherwise agreed.

Key features

- Includes core questions on organisation, people and governance plus options covering content moderation, automation and safety
- Incorporates the Internet Commission's 2019 Evaluation Framework for Content Moderation and draws on learning from its Accountability Report 1.0
- Aligns with key indicators in the Ranking Digital Rights framework, reflects the Voluntary Principles to Counter Online Child Sexual Exploitation and Abuse, the requirements of the ICO Age Appropriate Design Code and the implementation guide for the Council of Europe's Guidelines to respect, protect and fulfil the rights of the child in the digital environment: Recommendation CM/Rec(2018)7 of the Committee of Ministers
- Incorporates feedback from researchers, regulators and participating organisations
- Takes a European and international perspective, independent of government and industry.

Structure

The Framework invites answers to some core questions plus additional sections as relevant:

1. Organisation, people and governance: about the organisation's scope and purpose, the people concerned and its governance.
    a. Organisation: What does it do? Why, where and how?
    b. People: How does the organisation connect and interact with individuals?
    c. Governance: How does the organisation oversee digital trust, safety and freedom of expression?

Additional sections:

1. Content moderation: how is harmful and illegal contact, content, or conduct discovered and acted upon?
    a. Policies: the organisation's own rules, guidelines and procedures
    b. Reporting: How are you alerted to potential breaches of local laws and your own rules, for both harmful and illegal content?
    c. Moderation: How are decisions made to take action about content?
    d. Notice: How are flaggers and content creators notified?
    e. Appeals: How can decisions be challenged and what happens when they are?
    f. Resources: What human resources are applied to moderating content?

2. Automation: how are intelligent systems used to promote and/or moderate online content?
    a. Content promotion: behavioural targeting
    b. Automated content Moderation tools, in particular those based on Artificial Intelligence (AI) and/or Machine Learning (ML)

3. Safety: what measures are in place to protect people's health and well-being?

THE INTERNET COMMISSION

hello@inetco.org
www.inetco.org

+44 (0)20 8242 4066