Written evidence submitted by Trustpilot

Online Safety Bill - Trustpilot Submission

Summary

As a leading online reviews platform, we appreciate the opportunity to provide our perspective on the draft Online Safety Bill.

We welcome the UK government's aim to tackle illegal and harmful activities online, but we believe the current draft Bill risks creating onerous and ambiguous obligations that will be difficult for online services to meet. The Bill's multiple layers and duties of care that extend obligations to illegal and harmful content is compounded by a lack of detail on how to achieve compliance. At the same time, the Bill gives too little consideration as to how to effectively balance people's competing rights and interests.

To improve its clarity and effectiveness, and ensure better legal certainty for services, we suggest that the Bill should be amended to:

- narrow the scope of the "safety regime" and its preventative approach to a more practical compromise, including clearer requirements for services, limitations on the use of algorithms that proliferate illegal content, a less administrative-heavy form of risk mitigation and limiting the scope of take-down obligations for services to *illegal* content rather than regulating subjective "harmful content" which may not even give rise to liability for the content's publisher;
- include a more detailed consideration of how services can balance rights such as freedom of expression against the removal of harmful content in practice;
- provide clear definitions for the types of content in scope, taking into account that users of services will also need to be able to understand these enough to know what is permitted and when it is valid to challenge or dispute decisions;
- allow services to innovate for appropriate solutions that suit their models, such as for building safety into their designs and processes;
- avoid imposing "general duties", and sharpen the broad wording of Section 9 that could be read as implying general monitoring obligations;
- avoid diverging significantly from the European approach to allow the UK market and the services wishing to enter or participate in it to comply and stay competitive, to the benefit of UK consumers;
- avoid the imposition of a positive obligation for services to take down content, unless imposed by a direct order served by an independent regulator;
- take a holistic approach to the UK online safety regime that (building on the objectives set out in Chapter 5) includes increasing digital literacy, and accommodating practical and realistic measures to educate UK users about what is permitted online. Extend education to include reinforcing foundations for the tolerance of differing views within a free and democratic society, and help people to understand how to distinguish the unease of disagreement from actual harm.

Part A: The Trustpilot platform and our reasons for responding

Introduction to Trustpilot
Trustpilot is an independent consumer reviews platform with European roots and a strong UK presence (we're also listed on London Stock Exchange). We host user-generated reviews that connect consumers and businesses and enable collaboration between them.

Trustpilot is free for consumers to use and provides a place to share and discover reviews of businesses, websites and products, to help everyone make better and more informed choices. Through sharing their reviews and feedback on consumer experiences, Trustpilot users help promote confidence, transparency and trust. Our service also provides tools for businesses to help them collect and use customer feedback to establish and grow their credibility, reputation and branding. Our 'freemium' model for businesses allows them to use our basic services (including inviting and responding to reviewers) for free, or subscribe to receive enhanced paid services such as analytics, marketing widgets and social media integrations.

Since our launch in Denmark in 2007, Trustpilot has grown to a network of eight global locations – two of which are in the UK. Our site hosts over 120 million reviews and each month receives in excess of three million new reviews worldwide.

Reason for submitting evidence
Trustpilot is an Internet service that hosts user-generated content in the form of online reviews that are uploaded, shared and encountered by UK-based users. We are likely to fall within the scope of and be impacted by the UK's Online Safety Bill.

Like many internet-based services, we operate across multiple jurisdictions. We're keen to see the UK adopt a workable approach to online safety that allows us to continue to competitively offer our services in the UK to consumers and businesses, as we do in other markets within Europe and the US.

To help achieve this, we're interested in sharing our practical perspective as a leading online reviews platform – a perspective which we believe contrasts with the views of the platforms currently dominating the discussions, such as the social media giants and platforms with ad-based business models.

As an online reviews platform, we're fully aware of our responsibility to protect people from harmful and illegal content. But we don't operate in the same way, nor do we tend to see the same types, volume and rapid proliferation of harmful content, as social media platforms do. We also do not solely rely on AI/machine learning to moderate content, or use algorithms to amplify the dissemination of trending (and potentially harmful) content. Precisely because we're different from many of the Big Tech players, we're keen to make sure the conversation about regulation stays nuanced. We want to help achieve legislation that works across different platform models – not just for the largest social media platforms and the most dominant business models. We also want to ensure that efforts to address the issues

involved give adequate consideration to the myriad of ways in which different types of services can be affected, thereby minimising the extent of unintended legal, social and economic consequences.

As a two-sided service that sits between consumers and businesses, we already seek to balance competing fundamental rights such as freedom of expression and the freedom to do business. Therefore, we consider our perspective to be valuable in helping to inform the discussion around online harms and considerations for striking the right balance between people's different rights and freedoms.

Part B: Trustpilot response to DCMS Sub-Committee questions 1-6

1. **How has the shifting focus between 'online harms' and 'online safety' influenced the development of the new regime and draft Bill?**

A focus on online *safety* probably represents a wider approach than minimising online *harms*. "Online safety" is likely broader than "online harms" because it suggests protection, security or freedom from harm, injury, risk, danger, or damage. This preventative approach results in a draft Bill that is perhaps admirable in its lofty ambition to "set a global standard for safety online" but risks being unworkable in practice.

The Bill combines an ambitious goal with a lack of detail about the standards it seeks to set and sparse guidance on how services can realistically meet the responsibilities imposed by the multiple layers of broad duties. There is also little detail on, and it is therefore difficult to grasp, how services might minimise the risk of curbing people's fundamental rights. For example, how should services meet, and adequately document, the Section 12 "duty to have regard to the importance of protecting users' right to freedom of expression within the law"? This wording falls short of the substantive protection of users' rights and freedoms that is necessary. And where multiple users with competing and contradictory rights and freedoms are involved (e.g. in our case, consumers and businesses), it will be practically difficult for services to argue compliance in one direction or the other. What constitutes having "regard"? It is foreseeable that there may be considerable practical difficulty involved in maintaining documentation that demonstrates adherence to all of a service's broad duties, not to mention managing the added administrative burden and cost.

The draft Bill provides a broad framework, with details yet to be filled in by Ofcom and the Secretary of State via secondary legislation and codes of practice. It's therefore no small task to try and assess the scope, reach and extent of the potential obligations. For example, since regulations specifying "Category 1 threshold conditions" for regulated user-to-user services are yet to be set out, we cannot know with certainty whether and to what extent Trustpilot will have duties concerning content that is harmful to adults. As another example, Section 7(b) creates a duty to "keep an illegal content risk assessment up to date, including when Ofcom make any significant change to a risk profile that relates to services of the kind in question." How will Ofcom profile certain services, how frequently should the assessment be reviewed to stay "up to date", and what is a "significant change"? At this point, uncertainty weaves a consistent thread throughout the Bill.

The draft Bill doesn't clearly specify the types of content that should be within scope, nor does it contain easy-to-understand definitions and set out defensible processes for how to assess content (see question 2 below). This raises the question as to how the comprehensive responsibility placed on services can realistically be met in practice *and* risks opening up platforms to disputes – including from users who frequently have their own views about what should and should not be permissible online. Vague notions of safety underpinned by subjective definitions are more likely to appear to dissenting consumers as censorship cloaked as safety, and it is foreseeable that challenges to a service's subjective assessments on content will be difficult to decisively resolve without independent intervention. For example, when does a service have "reasonable grounds" to believe content constitutes a relevant offence? And how can this be documented to the satisfaction of a user who – as a starting point – may not even be aware that what is illegal offline is also illegal online, let alone that a platform only has to believe and *not prove* that the content infringes the law?

In the online reviews space, we experience that negative reviews written about a business can feel personal for business owners and provoke emotional responses to a greater degree than negative product reviews about a specific product. Negative comments about a business or its owner's actions can be seen as criticism. While it is clear (from at least Section 46(8)) in the Bill that "content that is harmful to adults" doesn't extend to financial impact or the way in which a service featured in the content may be performed, application of this might be less clear-cut to online reviews where reviewers frequently blend their opinions on services, products and expectations with stories about what happened, and business owners (especially small businesses or sole proprietors) increasingly blur the line between their personal lives, business efforts and their identity. Businesses who are negatively reviewed will want clear and adequate reasoning to explain why content they might perceive as "harmful" by their own definition does not necessarily qualify for removal under the UK's online safety regime.

Decisions on the likely level of harm for statements of opinion, including online reviews and replies to online reviews, will be difficult to adjudicate and defend. In our experience, feelings and emotions tend to quickly overwhelm taking a rational approach and people find it hard to distinguish between the idea that other people have the right to have a different opinion versus *supporting or accepting* that differing opinion. Where a business and a consumer see one transaction in two very different ways, they can be uneasy with accepting that there are two different viewpoints, or two different opinions about the same circumstances. If one side sees the other as "incorrect", the inevitable solution is seen to be the removal or silencing of the other's voice.

Most online platforms also operate across multiple jurisdictions. Yet even within Europe and the UK they encounter considerable cultural differences. For example, Trustpilot is headquartered in Denmark which tends to be liberal in views, tolerant of nudity, and generally preserves a high bar regarding freedom of speech (e.g. consider the 2006 Mohammed cartoons https://www.theatlantic.com/international/archive/2016/03/flemming-rose-danish-cartoons/473670/). However, these standards can be different in the UK. How should services assess content that might be allowed in Denmark but is offensive in the UK? Geo-blocking might be presumed as a solution, but this can be complex in practice and is

not necessarily easily implemented for services that have not originally been designed with this in mind. It could also be contrary to EU geo-blocking legislation.

The new era of online services and platforms (and not least social media) gives voice to people who may previously have been silenced, marginalised, or unable to access large audiences. At the same time, humans are curious by nature and tend towards – rather than away from – controversial online posts. The current proliferation of sensational and sometimes harmful or illegal content is well-documented, but the problem is also one that should be addressed from multiple angles, of which increasing the responsibility of platforms is only one such angle. Free speech is a fundamental right and a central precondition for a democratic and free society. It stands in contrast to the idea that people have a right to be protected from being insulted or offended. The UK's online safety regime should help people to accept that others can have a different opinion even if the content in question provokes a feeling of anger or disempowerment, and this is very different to and must be distinguished from certain types of harmful and illegal content that need to be taken down.

While imposing legitimate limitations (e.g. on the use of algorithms that proliferate illegal content) and requiring risk mitigation is not without merit, an all-encompassing approach to "safety" which requires services to prevent harm and effectively protect people (sometimes from themselves), ignores the practical difficulty of services being able to achieve this and carries with it the very real risk of infringing human rights and free speech, whether inadvertently in an honest application of the rules, or deliberately via misuse. It's difficult to see how negative impacts on human rights will not be one inevitable result *unless*, at the very minimum, such an approach also clearly articulates fundamental details on how to achieve it. Whether or not the overall aim is feasible at all – even with the help of clearly articulated guidance on how to apply subjective standards – is addressed further below.

2. **Is it necessary to have an explicit definition and process for determining harm to children and adults in the Online Safety Bill, and what should it be?**

Yes. If harms are to be included in and addressed by this Bill, then explicit definitions of the relevant harms, and clear processes for determining those harms to children and adults are imperative.

We reiterate here our view that harms should *not* be included; instead the Bill should be limited to addressing illegal content. This is because harmful content is always subjective, nuanced and hard to define and assess. Even if content moderation is conducted diligently, it carries the risk of either infringing consumer rights to freedom of expression and non-discrimination, or on the other hand – for a two-sided and open online reviews platform such as Trustpilot – curbing people's freedom to carry on a business. Including subjective harms within the Bill in fact risks undermining the very concept of open platforms such as Trustpilot, where consumers should be able to freely voice their views without undue censorship.

While the anticipated codes of practice may assist in setting out steps to comply, it is unclear how it will be possible for Ofcom to take into account the different types of services handling different types of content generated by different types of users, and viewed and

disseminated in different ways. This requires deep knowledge of the different types of platforms, how they work, and how people use them.

The current absence of clarity in the definitions and processes also opens up platforms to disputes from users who do not accept decisions on the removal or non-removal of content. Even if clear reasons for removal or non-removal are explained to users, the presence of a subjective assessment significantly escalates the likelihood of dispute and potentially, conflict. Where “harmful content” may not even give rise to liability for the content’s publisher (e.g. a user), how willingly will such a user accept that it must be removed from a service?

In our view, harms need to be defined in a way that is useful for both sides – i.e. for a user trying to understand why their content is removed, and a service trying to justify why their content moderator has removed it. The current definition encompasses content that the service provider has “reasonable grounds to believe” gives rise to a “material risk” of a “significant adverse physical or psychological impact on an [adult/child] of ordinary sensibilities”. While this might be further specified in future regulation, the mere fact that there are multiple elements to the definition will make it seem complex to the average user. Such a complex approach doesn’t eliminate the no-win situation that services will inevitably face when adjudicating subjective content. Vague terms such as “significant” will need to be explained. And terms such as “ordinary sensibilities” are less familiar and more difficult to explain/justify/defend to lay people (making up the majority of users) than, say, a measure such as “a reasonable person”. How will services explain and justify the removal (or non-removal) of content without seeming as if they are deliberately using confusing jargon? Rewording in clearer and simpler language is not only preferable to using separate regulation to further specify these definitions, it’s probably necessary to provide services with the ability to explain to users how they’re operating, and maintain those users’ trust.

3. **Does the draft Bill focus enough on the ways tech companies could be encouraged to consider safety and/or the risk of harm in platform design and the systems and processes that they put in place?**

Yes. In fact the requirements may go too far in scope, requiring for example “illegal content risk assessments before making *any significant change* to *any aspect of the design or operation* of a service to which such an assessment is relevant.”

This is a broad and vague requirement that might not be helpful in practice. Diligent platforms are already building safety by design into their processes, and given the variety of different service models, it’s usually the platforms themselves that are best-placed to decide the approach that works for them.

Platforms are increasingly incentivised to build trust with users and be transparent about the ways in which they are considering safety or risk of harm in the design of their services. The public’s increasing awareness of online harms and safety and growing skepticism of online services makes earning trust crucial for all tech companies, and the building of safety considerations into platform design and systems/processes a necessary undertaking for maintaining a user base. There is also considerable discourse and best practices available

on safety by design and a risk-based approach to platform design, and trust principles for services to draw on.

At Trustpilot, we're already doing work in this area, with plans to expand it via trust-driven projects. With respect to the inevitable tension between creativity and rationalisation, our preferred approach is not to work to a set of strict, standard check-the-box requirements, but to adapt our approach to new projects as required. A set of high-level principles to consider when making changes to products and services would work more effectively than dictating specific systems and processes which may not be relevant and could stifle innovation.

It is practically impossible for the legislation to consider all platform types and models and their permutations, and a single approach doesn't fit all. The Bill should not go any further than setting out the issues that need to be considered, and suggesting high-level principles. At most, it could be beneficial to create and promote a repository to share best practices. No further mandatory requirements should be added; there are unlikely to be benefits in doing so since the majority of platforms differ in their models, operation or features, and need the flexibility to adapt these. Further specification is more likely to be viewed as an administrative impediment and a diversion of resources rather than a useful part of the creative process that results in safer services.

4. **What are the key omissions to the draft Bill, such as a general safety duty or powers to deal with urgent security threats, and (how) could they be practically included without compromising rights such as freedom of expression?**

Key omissions include:

- substantive safeguards to preserve rights to freedom of expression and privacy: The Section 12 "duty to have regard to the importance of (a) protecting users' right to freedom of expression within the law, and (b) protecting users from unwarranted infringements of privacy, when deciding on, and implementing, safety policies and procedures" lacks substance. It offers little real insight into what needs to be done, or how services are to achieve it. While Category 1 services need to keep impact assessments up to date and specify protective steps taken, if a lesser standard is assumed for "All services", it is not clear what this might entail. It will also be difficult to demonstrate compliance with or defend;
- specificity in definitions: See answers to questions 1 and 2 above. In addition, definitions such as "of democratic importance" or "journalistic content" are only broadly outlined. It seems unlikely – but isn't clear – whether online review content could be considered to fall within either of these categories; and
- an *outline* of the criteria that will be used to determine how platforms will be categorised: The Bill provides that the Secretary of State will decide how a platform is categorised, but there are no criteria, nor any useful detail that will help platforms to anticipate their obligations. Given the diversity of platform models, this lack of transparency about which factors will be involved in categorisation could increase the risk that services are miscategorised as Category 1 and required to meet the most stringent transparency and accountability standards. Such processes require a significant amount of resources and labour. And platforms that have large user bases might be presumed to be large in size, have considerable influence, and have a

corresponding high level of resources at their disposal, but *may not necessarily* do so. This is especially the case if they are open platforms, since these services may have a large number of viewers of their content, but not necessarily a high number of regular users of their services. These two different ways of measuring users can produce significantly different assessments of the size and influence of a service. A further important distinction in considering the nature of services is the business model: those driven by ad-based revenue streams or the sale of people's personal data tend to operate differently to software-as-a-service platforms such as Trustpilot.

Platforms that are miscategorised as larger platforms may not have the resources to meet more requirements or pay the corresponding fines, putting them at a significant disadvantage. The UK government should therefore provide greater clarity around how platforms would be categorised for the purposes of the draft Bill, to allow adequate time for preparation (which may include hiring and upskilling) and provide companies sufficient time to adapt to the new responsibilities and test and adopt appropriate technical solutions.

We do not view a general safety duty as an 'omission', since:

- this is likely impossible to achieve in practice without effectively imposing a requirement for general monitoring of content; and
- the current Bill could be read to already imply such a general duty.

Taking these points in turn: first, compliance with a general safety duty would inevitably require vigilance in content present on the service. Maintaining it would be resource-intensive and a barrier to entry for SMEs, which could potentially result in reduced competition in the market by stifling innovation from smaller and less established platforms without sufficient resources and/or technical capability and who would be unable to absorb the costs involved in maintaining such a high level of vigilance or take the risk of liability for such content. It would also increase the risk of eroding people's right to freedom of expression as user content would be monitored and restricted, undermining the concept of 'open platforms' such as Trustpilot (which allow consumers and businesses to collaborate and freely express themselves) and forcing platforms to adopt more 'closed' models. This would potentially infringe free speech and the dissemination of user ideas, as well as the freedom of platforms to conduct their own business. This seems an unlikely aim for UK legislation.

Second, the current Bill includes a broad safety duty covering illegal content, which could operate in practice similarly to a general safety duty. Section 9 applies to all service providers and imposes at (2): *"A duty [...] to take proportionate steps to mitigate and effectively manage the risks of harm to individuals, as identified in the most recent illegal content risk assessment [...]"* and at (3): *"A duty to operate a service using proportionate systems and processes designed to"* minimise the presence of and length of time for which priority illegal content is present, minimise the dissemination of priority illegal content, and swiftly take down such content on being made aware of it.

5. **Are there any contested inclusions, tensions or contradictions in the draft Bill that need to be more carefully considered before the final Bill is put to Parliament?**

Yes. As outlined above, more careful consideration needs to be given to the protection of fundamental rights such as freedom of speech, and how to achieve sufficient legal certainty for services. Alongside this, it's not yet clear how services can deliver safe solutions to meet all of the legal requirements; for example, the technology for gating platforms to keep children off them is still developing and improving.

6. **What are the lessons that the Government should learn when directly comparing the draft Bill to existing and proposed legislation around the world?**

Platforms typically operate across multiple jurisdictions. Diverging the UK's approach too much from Europe will make it very difficult for platforms to operate in the UK, which may result in a lack of innovation and entry into the UK market. Considerable divergence can already be seen between the draft UK Bill and the EU Commission's draft Digital Services Act (DSA). For example, the difficulties of legislating subjective categories of content via a systems and processes approach are evidenced by the EU's decision to avoid precisely this; the DSA takes a risk-based approach, but limits notice-and-take-down requirements to illegal content.

Another key difference is that the UK's draft Online Safety Bill creates a *positive* obligation to take down content while the EU's DSA takes a different approach to liability by removing the safe harbour protection afforded to services for non-compliance with take-down requirements.

Further, outright inconsistencies between UK and EU rules will be detrimental to competition. For example, if Section 9(3)(a) through (c) can be read as effectively implying a general monitoring obligation, then such an approach is inconsistent with the EU's draft Digital Services Act. This not only raises concerns about the resources that services will need to devote to comply, it also invites questions regarding the quality of such checking of content, costs (as anticipated by the Online Safety Bill - Impact Assessment at 166) and importantly, the extent to which it endangers freedom of expression because it must be imposed at scale. While "…it is expected that undertaking additional content moderation (through hiring additional content moderators or using automated moderation) will represent the largest compliance cost faced by in-scope businesses" (Impact Assessment [166]), these two "solutions" are by no means equal alternatives in terms of cost or quality. Services that decide to use more artificial intelligence to identify harmful content invite machine-lead dangers in the form of inaccurate take-downs, wrongly censored language (especially irony, sarcasm and satire in the UK), as well as racial bias.

Other regimes that address "safety", such as Australia's Online Safety Act, impose rules relating to individual pieces of content. This balances the ambiguity of a broad approach with more specific details on how to take down or disable access to individual pieces of content quickly. In contrast, the UK's approach combines a broad "safety" approach with a broad focus on systems and processes. This is likely to be difficult for services to comply with.

Finally, the UK should learn from European jurisdictions such as France and Germany to avoid extensive opposition and potential legal challenges to legislative regimes seen as

going too far in their obligations for tackling online content. For example, the key provisions on content removal contained in France's Avia law were limited to illegal content, but were struck down as jeopardising freedom of expression and communication rights due to several factors, including the threat of large fines for non-compliance, the short time-frame for removal and the difficulties in assessing hate content quickly. The German Act to Improve Enforcement of the Law in Social Networks (NetzDG) from 2018 continues to face widespread criticism for encroaching on freedom of expression.