Tech Notices Webinar Transcript

23 January 2025, 03:04pm

That today, and they might help you sort of guide you in posing questions to us using that Q&A function. So the first question we had was if a web shop allows customer reviews on their website with this be considered a user to user service. So I'd say this question goes to, well, who's in scope of the Online Safety Act? That's a much broader question than what we're going to be looking at today where we're focusing specifically on the tech notices powers.

But I would like to draw your attention to our digital support service that's hosted on the Ofcom website and includes a range of tools to help you navigate the online safety framework. One of those tools is our regulations checker and that can help you check whether your service is likely to be covered by the act. Now, I say likely to because ultimately it's for services to assess themselves whether they're subject to the regulations.But hopefully these tools can point you in the right direction, and I think we can put a link to those resources.Somewhere, possibly in the Q&A, as the chat is down.

Secondly, question was what will the impact of your work on 3D printed firearms be for the online acquisition of blueprints and making instructions? So I'd say there are sort of two answers to this. One, the broad answer is that to the extent that those blueprints are illegal content and especially if they're priority illegal content as defined in the act, then we'd expect our wider work on illegal harms, like the statement and the codes of practise we published in December to have an impact on that area.

But for the purposes of today, as you're going to see our tech notices powers specifically relate to terrorism and child sexual exploitation and abuse content only. So I just wanted to flag this question. I think it's helpful to keep that in mind in terms

of the scope of what we're talking about now. I do know that of at least one conviction recently in the UK of an individual for collecting blueprints of three 3D printed firearms. And that was where the court found that that activity was a breach of various terrorism offences.

So I guess I'd say in principle, if blueprints are illegal content, that is also terrorism content, then that is something that could be impacted by the way that we're talking about today. I won't go into any more detail on that now. Hopefully whoever asked that question, you'll get a bit more of a sense from the presentation of what the impact could be. But I'd also caution that it's probably a little too early in the development of this framework to say what the impacts will be.So that's a couple of questions. I'm looking forward to sort of answering some more questions in due course.

Now just quickly back to housekeeping, you may also notice that we're recording this webinar, but this is for internal purposes only and I say we're only recording the presentation section, not the Q&A. Technically, if you experience issues with the stream, we'd recommend leaving and rejoining. So basically turn it off and on.
And then finally, just picking up on a point that's come up when we talk about this work in other contexts, people sometimes ask us if there's likely to be anything triggering in the session.

I can confirm there's nothing in the way of triggering content directly included on materials for today, but the underlying harms that this power relates to mainly terrorism and csea can be really upsetting. So just wanted to make you aware in advance that there could potentially be discussion of, for example, the way homes like that manifest in online spaces. I just wanted to mention that. So you're prepared in advance in case those kind of topics come up in discussion.So that's everything I wanted to say by way of introduction and housekeeping.And now I'll hand over to Alex for the presentation.

Thank you, Andrew.So first of all, contents, what are we going to go through today in our presentation? So firstly, we're going to speak to the powers and deliverables. Under this part, Section 121 of the Online Safety Act will then speak to the guidance for providers how often proposals to exercise its technology notice functions well,

then address our policy proposals for setting minimum standards of accuracy, which I'll I'll explain what that is in just a second, and then we'll move to our question and answer portion of the session.

So we're going to start with an overview of the technology. Notice powers, the Online Safety Act enables Ofcom to make a provider use, or in some cases develop a specific accredited technology to tackle terrorism or child sexual exploitation or abuse content on their sites and apps. And these actions would be taken through issuing a technology notice. And it's worth noting that on a user to user service which is a service enabling communications between users. And in relation to csea content specifically, the power can be used to require technologies to identify content communicated both publicly and privately. But before we can use this power, there are a few important steps that have to be taken. So first of all, the act requires Ofcom to provide advice to the Secretary of State on how to set minimum standards of accuracy against which suitable technologies for the purpose of a notice can then be accredited. The Act also requires Ofcom to issue guidance for providers regarding how we intend to use the powers and following advice from Ofcom, the decent Secretary of State will have to approve or public approve and publish minimum standards of accuracy then Ofcom or a person appointed by us will need to accredit technologies as meeting minimum standards of accuracy, so only once all of these steps have been taken can Ofcom then consider the use of this power.

So what we what we've done so far?

And while in mid-december we published our consultation on two of the elements that underpin this power, that was the the policy proposals for setting the minimum standards of accuracy ahead of issuing any advice to the Secretary of State as well as draught guidance to providers and how we proposed to exercise our techno dysfunctions.And this month, we'll also publish the 1st Annual report on the use of these powers. And this is just an annual reporting function about technologies which meet or are in the process of being developed to meet the minimum standards of accuracy.And Ofcom's exercise of its function in this area in the previous year.
The reason then for hosting this webinar today is to present an overview of the policy proposals outlined in the consultation and just answer any questions that you might have on these. So if that's the power, how would Ofcom issue a technology

notice on the slide? We walk through what providers of part three services, which are the regulated services which could be required to to use an accredited technology or developer source. One can expect if Ofcom are considering issuing a technology notice, there's a few distinct phases here. So firstly, the initial assessment when Ofcom becomes aware of an issue. Will carry out an initial assessment to decide what action, if any. It would be appropriate to take and. It's important to note that that Ofcom has a variety of tools at our disposal to attempt to resolve issues and the technology notice power is just one of those.

Secondly, information gathering at this stage Ofcom may use our information gathering powers under the Act or via other methods available to us.
One such information gathering power is to obtain a skilled person's report, and this is a report that often must, as according to the Act obtained before we issue a technology notice.And the purpose of of this report is to assist in deciding whether to give a notice and advise on the requirements that might be imposed by such a notice.The relevant matters we would ask the skilled person to advise on will depend on the specific circumstances, but we may, for example, request that the school person's report explains the provider's existing systems and processes in place to deal with target content and how or or where accredited technology could be implemented alongside this.We then have the the warning notice where Ofcom and 10, so we're Ofcom intends to to issue technology notice and we will give a warning 1st and explaining our intention and we can only issue one of these warning warning notices after we've obtained a skilled person's report.And within the the warning. Notice this will include a summary of the skilled persons report and details around the requirements we intend to impose.There will then be the opportunity for services to make represent representations in response to that warning notice.
And in our draught guidance we've suggested, typically giving at least 20 working days for representations, although of course this is something that we're consulting on.We then have this final step on on the screen which is following the assessment of the representations Ofcom deciding whether or not to issue a technology notice, where we consider it necessary and fortunate to do so.And the time scales for compliance with the notice will be specified in the the notice itself, something that's not noted on the screen here, but it's also worth being aware of is the Ofcom also must carry out a review of services compliance with the technology notice and will

notify a service of of any outcome of that review.

So. So now a kind of integral part of this power is how Ofcom will take into account necessity and proportionality considerations in order to determine whether to issue a notice.This and this is important because this power can be used when Ofcom consider it necessary and proportionate and does not necessarily require a service to be in breach of their duties under the Act. So in section three of our guidance, we outline Ofcom's approach to considerations of necessity and proportionality when considering whether to issue a notice and on the slide here we have some examples of the sorts of considerations Ofcom have to take into account as per the act and we refer to these in the in our guidance as spec. They're loosely gripped on the screen here, so in the blue on the left here, the left three columns we have contextual risk based considerations and these include service type harm or content type and user base. And in purple the penultimate 2 columns from in from the right there those are human impacts, human rights impacts, and these include freedom of expression, privacy, data protection and journalistic freedoms. And then finally.
The final column on the right there in Orange.And it's just an important consideration. Will always be at what other tools does Ofcom have at our disposal? And again, that's just a reminder that the technology notice power is just one of many powers Ofcom have to deal with target content. We've also outlined some further considerations that Ofcom would have to take into account on a case by case in our guidance. And so for example, the cost of implementing a technology into a service is something that we we think would be important to take into account the technical feasibility of the service provider doing what would be required of them in the notice is is something we would consider and also we might consider whether compatibility testing is appropriate.To inform our view that the compatibility testing of the technology and the service.And then the also the potential impact of the technology notice in, in actually reducing the amount of terrorism and CSA content for example.That's quite a high level view of what services subject to a technology notice can expect from the process of issuing a notice, as well as how often kind of will consider those necessity and proportionality considerations before issuing a notice. But there are further details on all of this in our draught guidance.
I'm gonna move us on now, though, to our policy proposals for setting minimum standards of accuracy and which we would accredit technologies against.
I think first and foremost, it's it's worth clarifying what's in scope here.

So. So what what is accredited technology and the Act requires technologies to be accredited before Ofcom can require their use as part of one of these tech notices. And this means that technologies that could be accredited include the technologies capable, any technology capable of detecting terrorism under CSA content, which is the target content of this power. And then the specific technique, specific technological products, tools or solutions? And I think that that's quite an important distinction here from other areas of the act where we speak kind of in broader terms about technologies, whereas here we're targeting really specific product level technologies and of course as per the wording of the app, technologies must meet those minimum standards of accuracy to become accredited. So you'll notice throughout this presentation and our consultation that we discussed both the minimum standards of accuracy and accreditation as inevitably intertwined. And that's because in order to set a standard, we do have to have to understand the assessment framework. In our consultation, we've proposed an audit based assessment framework with a supplementary optional second stage for lab based independent performance testing.

So the audit based assessment that is akin to an audit model.Tech developers would provide evidence of their technology meeting specific objectives, and this evidence would then be independently scored by Ofcom or persons nominated by Ofcom. And and then independent performance testing, which as I say is an optional second stage which we're consulting on that's a lab based proposal and where technologies would be categorised and tested against data sets to independently compare their performance against comparable technologies. And we're proposing that performance thresholds. There are sets based on the top performing technologies in each category.So diving a little bit deeper into our proposals, first, if we look at the audit based assessment. We've developed the audit base assessment around 4 principles. Those are technical, performance, fairness, robustness and maintainability, and we've developed this assessment to apply to all in scope technologies, regardless of the data input or harm assessed and our proposal sets out objectives under each one of these principles. Against which the tech provider would be expected to provide evidence and then score sufficiently well against.
So if we take robustness as an example, you'll see on the slide that we've got the list of objectives that we've consulted on.The next step then, is to determine how does a tech developer evidence that they've met these objectives, and and we've developed

a set of questions and required evidence that we think would answer these questions, which would then be scored by the accreditor. So if we take that final objective there, which is detection and mitigation of threats.

On this slide, we've we've got an example question. So which? So example question and this could be one of many questions that could be used to assess this objective, but this is just one example for detection and mitigation of threats. So the example question here is how does the technology perform when subjected to input perturbation attacks such as adding noise, altering colours or modifying words? The technology developer would be expected to submit evidence which we could then score 01 or five points. And we've designed the scoring system to be imbalanced, to put greater weight on robust evidence. So we're just walking you through that. A score of 0 would be awarded for no or extremely limited evidence. A score of one would be awarded for limited evidence, which perhaps includes some perturbation tests, records of limited resilience outcomes, or or partial analysis of results.And then a score of five would be awarded for what we consider extensive evidence. So example documents we're looking for here would include detailed records of resilience across multiple scenarios, continuous monitoring reports, evidence of sustained robustness, detailed documentations of tests on various data transformations, things like this. But if we just zoom back out to look at how this, then all fits together and is scored to achieve the minimum standards of accuracy. So here you can see each of those four principles on the screen.
With their objective listed on objectives listed underneath and using the evidence submitted and each objective will be scored and then normalised under each principle to a score out of 100. So you'll have one score out of 100 for each one of those principal 4 principles. We're calling this the principal score and proposing that the technology must achieve a minimum score of 50 out of 100 on each principal. With the principal score is then and these are added up with maintainability, weighted slightly lower just to acknowledge that reggregation periods.
Of four years, which is what we've proposed in our consultation, mitigates this to some extent. And then again the score would be normalised to a total possible score of 100. So you then have one overall score and we propose that the the technology must achieve a minimum score of 60 out of 100 overall to meet the minimum standards of accuracy.

So what does all of this mean for our advices, sector of state and the actual minimum standard of accuracy? So we are proposing that the Secretary of State would publish the principles and objectives, the details of the scoring framework, the Ofcom have proposed and the final score thresholds to meet the minimum standards.
Just as a reminder there that the principal score was 58100 and then the overall score would be 60 out of 100.So we now have a minimum standard of accuracy for the audit based assessment. And to be clear, in our in our consultation, we're proposing to do this audit based assessment 1st and in every case, but we do also have in our consultation a second optional stage based on supplementary independent performance testing.

So if we take a closer look at this, this part of our proposal and it's a proposal not in quite a few parts. So if I just start with the basics. Technologies would be tested in specific categories based on the harm type and data type addressed. And so for example, see some images terrorism URLs in these categories we've proposed to test against the F1 score and the metrics listed on the screen here. And in order to be accredited, we would expect technologies to meet or exceed a set threshold on the F1 score and at least one other metric on this list. But I think then kind of a natural next question is, but what are these thresholds? In our consultation, we've proposed to use benchmarked thresholds which would be calculated, published and periodically updated by Ofcom using a mechanism approved and published by the Secretary of State. So in other words, the threshold will be set based on perform best performing technologies in each category. Importantly, this won't be a prescribed threshold such as 75% precision or 95% recall. Based on predetermined performance levels deemed acceptable for the illegal content in question, which would not necessarily reflect the current capabilities of technology submitted for accreditation, we've explained why we've come to this conclusion in our consultation. But among the reasons for opting for the benchmark thresholds includes the the thresholds would be based on performance of similar technologies under identical conditions using the same tests, metrics and data. And they offer a realistic reflection of the current capabilities of available technologies in the market. They also then act as a reference point for improvements in innovation, which we think makes this quite an agile approach which can adapt to market changes in the market, technical capabilities and also that evolving harms landscape. And importantly, it means that

we are able to acquire at the best technologies out there and be ready to use the technology notice powers.

With benchmarking. There are a few ways we could approach this, and we've consulted on two options, so mechanism A and this would be set at the 75th percentile of all submitted technologies during the previous testing period. I'll explain what a testing period is on the next slide, but bear with me and mechanism B and would be set at the 90th percentile of the single top performing technology of the previous testing period.But what are periods? So an important part of this proposal is this idea of the previous testing period.And essentially all this means is a set period of time which a threshold applies, and before it's updated. So in our consultation we've proposed that the benchmark threshold be updated every four years.This graphic attempts to illustrate this, and I can walk you through an example. So let's jump to period 2, which is the second green band in this graphic.
On the right here.So Threshold 2 is published at the start of this period. The threshold's then being calculated based on the performance of the technologies tested in the previous period one.And we've also included some examples of of testing during the threshold periods, just kind of at the bottom of that graphic there. And this is just to illustrate that no matter when the technology achieves accredited status and it's accredited to the end of the period. So for example, if you're accredited 2 years into a four year period, you were then only accredited for two the two years till the end of that period.And you'll note that the initial submission period is a little different. That is, the far left of this graphic it where we have the purple band.And that's just because there's there. There won't be a threshold. Yeah, at that stage. And this is the only accreditation period during which we wouldn't have a prediso threshold. And that's just why we operationalize the accreditation scheme.

So what does all of this mean? For our advice, the Secretary of State and the actual minimum standards of accuracy as a reminder, this is an optional second stage that we're consulting on and what we would propose is to advise the Secretary of State to include, within the minimum standards of accuracy the require the requirement to to. Pass both the F1 score and at least one of the other metrics from an UN agreed list of metrics and prescribe a mechanism for Ofcom to use to calculate benchmarks. Needed to meet the minimum standards of accuracy.So putting this back into context of the broader accreditation picture and Minimum Standard Act of accuracy,

we now have a standard for an audit based assessment, including the principles and objectives, the scoring framework and the the score thresholds.As well as that second optional stage of independent performance testing where we have our threshold set at a percentile of tested technologies and then the technologies having to meet or exceed F1 score in at least one other metric.

Zooming out just one step further and we can also now see how the design of these proposals acts as a as a funnelling effect, so applicants will be asked to apply for accreditation using a basic screening application form and which determines quite basic information such as the nature of the technology, the output it produces, the content it targets, basically just essential information for for Ofcom to understand whether the tech is is in scope for accreditation.We then, as part of this funnel, have the audit based assessment and then the option for the independent performance testing. This then gets us to that tip of the funnel and at this point Ofcom will have gathered pools or categories of accredited technologies categorised by the purpose of the technology.And the output then is that Ofcom are able to use the technology notice power where necessary and proportionate and and then at this stage there's option to do further kind of compatibility testing. Once we have further information about the conditions of a potential notice. So. The sorts of information there includes, you know, the harm type, the data type, the platform in question, et cetera.