Content Moderation – Adopting a content moderation programme for the digital age.

[00:00:00.250] - Introduction

Rob Corbett, Partner and Head of the Technology and Innovation Group, chairs this panel discussion on content moderation. The topics include key changes for content moderation programmes, considering AI in content moderation, new rules for content moderation and content removal and governance obligations under the Digital Services Act.

[00:00:21.110] - Rob Corbett, Partner

So I'm going to be joined for our next session by Olivia, who have met earlier, Ciara Anderson and Rosemarie Blake and we're going to have a chat about content moderation, adopting a content moderation programme for the digital age. So, as we've seen through each of the discussions we've had this morning, the rules as to what constitutes illegal or harmful content are changing, the obligations on those who are hosts of online content, whether it's illegal or harmful, are also changing and it sounds from the last panel like there's a serious enough lift required, particularly of those operators on the upper end of the inverse pyramid and they got to get it right by the 17 February, because the DSA is already in partial force, but it will be in full force by then. So this panel, I suppose we're going to look at some of the sort of pragmatic stuff that that means. I know some people in the room are in the content moderation area, some of you aren't, but may need to be in due course. So many of those on the hosting services will already have content moderation programmes. They're made up, I suppose, on the front end of a combination of terms and conditions, policies, community standards, so a sort of a self-regulation,

[00:01:37.210] - Rob Corbett, Partner

sometimes companies will adhere to codes of conduct on illegal hate speech or codes of practise on disinformation and that type of thing and then there's the legal side of things, the things that you're currently already required to take down when you become aware of them and we've talked earlier about issues like CSAM material, the child sexual abuse material, intellectual property infringement, that type of thing and then with the OSMRA, what we're learning is that there will be a process required for the service providers to identify content that's harmful and to have a process in place to address it, including with the risk assessments and so forth. So a lot to digest in there. So maybe I'll kick off with a question for Olivia, I think, so you set the scene earlier, Olivia, in terms of the whole broad range of regulatory changes that are coming down the track for intermediaries and then we heard from Commissioner Hodnett and Commissioner Dixon about how they're going to work together as regulators and upwards with their European Union colleagues in respect of the OSMRA, the DSA and the GDPR. So it strikes me that in order to comply with the OSMRA and the DSA, there'll be significant implications for content moderation teams in intermediation, but also for those seeking to have harmful or illegal content taken down.

[00:03:02.170] - Rob Corbett, Partner

So when I put on my parent's hat and think of some of the stuff I'd like to see removed from the Internet, I have a different perspective than when I'm an online content host being requested to take it down so maybe you'd help step us through some of the key changes we should be preparing for.

[00:03:17.630] - Olivia Mullooly, Partner

Sure, so just the first point in relation to the E-commerce Regime, so I noticed earlier that this was a notice and take down regime where you weren't liable for content until such time as you were put a notice of it and you had to then act expeditiously to remove it. There was never an obligation to do active monitoring of what was on your platform or your service, and that principle does remain intact. What they have clarified in the DSA is that if you are an intermediary service provider and you want to voluntarily explore what's on your system and you want to detect illegal content without necessarily waiting for somebody to tell you what's up there is that that won't affect your liability exemption under safe harbour, because that was never quite clear. There was always a sense of, if you start poking around and having a look at what's on your platform, are you putting yourself on notice? Are you affecting your safe harbour exemption? There's very much encouragement to conduct voluntary content moderation on your platform without affecting your safe harbour exemption. So that's if you do it on a voluntary basis, if you're an intermediary service provider that isn't subject to, as we said before, you start to move up the pyramid, when you're an online hosting provider, an online platform, particularly when you're an online platform, you do have to have a content moderation system and you have to make all of the information about how it works.

[00:04:41.740] - Olivia Mullooly, Partner

What kinds of notices are you providing? How many actions you've taken on foot of that? What is your content moderation made up of? You mentioned kind of previously, it was kind of a community guideline, standards based system terms and conditions, and plus the mix of kind of the legal obligations that take stuff down. The platform that need to explain now what their content moderation is made up of. Is it on the basis of Ts & Cs? Is it on the basis of illegal content, harmful content? And so forth. So they need to start by having a place on their website which isn't behind a login wall, because it has to be easily accessible, whereby all of this information has been made available as to how content moderation works, any use of AI, your complaint system, and how you've balanced kind of rights of individuals and all of that. The second thing that I suppose, and this is more relevant to the VLOPs and VLOSEs, is it must reflect your risk assessment. So if you're a VLOP or your VLOSEs, you have to actually take specific measures to identify illegal content on your platform, as opposed to if you're not a VLOPs and your VLOSEs, it's a notice and action process that you need to engage in

[00:05:52.940] - Olivia Mullooly, Partner

but for all of those online platforms, you need to look at your policies, your procedures, see are they fit for purpose? Are you set up to deal with the timings that were discussed earlier and that we will probably

discuss again on this panel, in terms of taking action in a timely manner in response to a notice to take down illegal content, making sure that the statement to the user who put it up there is provided and it has all the required information, you have a complaints mechanism, you have a system for dealing with this out of court dispute resolution and also bearing in mind that you have to take account of everyone's rights here in content moderation, you have the rights of the person who is complaining about the content, even the rights of the person who's put the content up there. This is new, like, we have a system here whereby fundamental rights, with some exceptions, were never supposed to be enforced by private entities and the balancing test was always to be done by governments and courts and state agencies and state actors rather than private entities. We now have a situation where private entities are now trying to call it in terms of how do we balance the fundamental rights on the platform and provide information on that.

[00:07:03.420] - Olivia Mullooly, Partner

That's really tricky. I mean, we have decades of case law on this where the balancing test between the right of freedom of expression versus the right to privacy, right to a good name, right to earn a livelihood, like the list goes on, constantly trying to get that balance right so that in itself is a tricky place to be and if you don't get it right, there's civil remedies that we'll come on to in our next session for people who can claim for compensation because you haven't got that balance right, and they are claiming that you didn't take account of their fundamental rights, so you know, you can see immediately there's an awful lot of complexity in that topic.

[00:07:38.530] - Rob Corbett, Partner

Yeah and I mean, even today, maybe, Rose, I'll bring you in here. I mean, the ability to spot content that's either illegal or harmful, or in breach of community guidelines or terms and conditions, it's an impossible task. It's a sisyphean task, I remember somebody, one of the platforms using that phrase in relation to even identifying the content and then having the systems in place to take it down. So in this respect, I suppose AI is your friend but then the deployment of AI in the context of content moderation itself brings particular and emerging challenges. So, Rose, is there anything else we need to consider when deploying AI in this type of an environment?

[00:08:16.110] - Rosemarie Blake, Senior Associate

Yeah Rob, there's lots to consider. So the DSA talks about algorithmic transparency. So for AI and content moderation broadly, the key things are, one, that you're providing transparency to the users of the platform. Two, transparency from a public reporting perspective and Lorraine touched on that earlier, mentioning the need to disclose the use of automated means in your public reports and then thirdly, the additional transparency and risk mitigation requirements applicable to VLOPs and VLOSEs so when you're looking at the transparency to users under the DSA, you have to provide information in the terms and conditions on the measures and tools for content moderation and this mentions explicitly the need to provide information on the use of AI and the human review aspect. So the level of detail on that provision of information isn't quite specified but we can be guided a bit more by some of the mentions in the

recitals, which say that you've got to be clear to the platform user on two key things. One, how the algorithms impact and influence the services that are being provided, and two, how the services are restricted through content moderation using algorithmic decision making. So secondly, looking at the transparency in the public reports, there's a requirement in relation to the transparency reporting of AI and content moderation, where you've got to be able to describe in a quite detailed way any use made of automated means for the purpose of content moderation, including how accurate that AI is, the rate of error and the safeguards applied to it.

[00:09:43.210] - Rosemarie Blake, Senior Associate

So to Olivia's point, in that balancing of the fundamental rights, you've got to be able to demonstrate that your use of AI isn't going to result in an unjustifiable removal of legal content, and that there are safeguards in place for preventing that occurring. So then, looking at the transparency and the additional pieces around VLOPs and VLOSEs. So there's an additional piece here on analysing systemic risks stemming from the platform and service function, including the systemic risks arising from the functioning of algorithms in the platform system. So last month, the European Centre for Algorithmic Transparency was launched, which is an AI research hub which will support the Commission in assessing compliance with these risk management obligations. Basically to identify the smoking guns to drive enforcement in the DSA, and hold some accountability there and test that risk management claims in the use of AI is going to stack up and then lastly, looking at the AI act and looking at the DSA's interaction with that. So it's really important to note here the draft AI Act is primarily going to relate to high risk systems so AI that's used, that poses a fundamental risk to people's fundamental rights and safety.

[00:10:49.880] - Rosemarie Blake, Senior Associate

The dial continues to move on that, but AI systems for content moderation are currently not likely to fall into that category. When you're using your AI in your content moderation system, you'll still be subject to transparency requirements under the draft AI Act and there'll be a focus on nondiscrimination. So that goes back to the safeguards point, you've got to be able to demonstrate you've got a diverse and broad data set for your algorithm to work effectively and in an unbiased way. So the kind of key takeaways there are transparency to users, being able to explain how the AI impacts the service they're interacting with, both from a presentation of the service and a restriction aspect, being able to explain the logic and design of the AI systems that are deployed in content moderation programmes in a meaningful way to comply with your transparency reporting obligations and then that systemic risk point to also note that you have an obligation to look at your risk mitigation obligations when using your AI.

[00:11:41.890] - Rob Corbett, Partner

Great. The platforms are going to be playing with fire here, right? I mean, the use of AI as you're talking there, ultimately potentially to profile people as criminals if they're deliberately posting illegal content, brings you right back into the GDPR world of profiling and human intervention and then the safeguards and the transparency and the risk assessments and so forth. I mean, the platforms may well be foisted on

their own patard by reference to the fact that they knew about the risks and they didn't or did they or did they not sufficiently address those risks? So we're learning a lot about not just the transparency piece, but also the backend operational points to comply with the DSA. So I don't know if there are any Olivia to you, maybe any sort of top tips, if that's what's coming over the horizon, what should the guys be thinking about in the room if they're involved in content moderation programmes?

[00:12:37.810] - Olivia Mullooly, Partner

Well, the first thing you'd have to do is decide where in the pyramid you sit and what action, if any, you need to take. Then you needed to have an assessment of which of the pieces of legislation that we picked up on earlier are relevant to your content moderation programme, because there's different steps and actions that need to take, because that will inform what response mechanism you need to have in place. There are certain kind of obligations to take action without undue delay under the DSA. There's a concept in the Online Safety and Media Regulation Act around taking action in relation to harmful online content. I think after, like two days, then the person can complain to the online safety commissioner. So you have all of these overlapping obligations, and the regulators mentioned earlier that they'll pick whichever one makes sense, but you still have to have the mechanisms in place to comply with them all. So doing an audit to see where do you sit in relation to that is going to be important. There are certain escalation triggers as well under the DSA, where you need to keep records so that you are understanding, like, is there anyone who is manifestly or frequently providing manifestly legal content because you need to take measures to deal with them.

[00:13:52.410] - Olivia Mullooly, Partner

If there is a suspicion of a criminal offence occurring on the platform, that needs to be escalated to law enforcement. There's various other issues, including in relation to your internal compliance, making sure that your content moderation is properly plugged into your compliance function. It's capturing the data that you need then to inform your risk assessments and any other steps that need to take so there is quite a lot that rides on your content moderation programme and don't forget that while AI is important and necessary and you need to explain how you're using it, there are restrictions in terms of where it can't be used in content moderation, and you will have to retain that human element in terms of like, for example, in relation to appeals. There's a lot of employment issues in terms of having a human content moderation team, in terms of risks assessing risks of personal injury. Are they suited to this particular type of work? So there's a big HR element in that that can't be forgotten about and employment law advice is frequently needed when you're staffing and resourcing and training members of that team. So those are just a few points.

[00:15:00.060] - Rob Corbett, Partner

Yeah, we're dealing with a stack of those personal injuries claims bought by content moderators who claim to have suffered post traumatic stress disorder and similar symptoms having been exposed to. So maybe, Ciara, I might turn to mean. So after we've determined what constitutes illegal content or content

that breaches the organisation's terms and conditions or its policies, the DSA at least sets out specific new rules that will apply across the EU for content moderation programmes so is this level of harmonisation, do you think it'll be welcomed by the platforms?

[00:15:30.890] - Ciara Anderson, Senior Associate

Absolutely. So, you know, I think, going back to first principles, that liability position really isn't changing. I think that's really important. If you're actually aware, you will be liable for content if you don't expeditiously remove it. What the DSA does is it provides a bit more colour to when you're actually aware and a procedure so ultimately, many organisations currently have content moderation programmes, but they're largely based outside of certain areas, like terrorist content, on their Ts and Cs, and on their own programmes that they've developed themselves. So this procedure is helpful so, for example, user notices in relation to legal content, they need to set out specifically the URL of the content and they need to provide an explanation that would allow a diligent provider to understand that the content is illegal without a detailed legal examination. So that is a bit of a standard which is helpful. Ultimately, if you receive a notice and it's not clear, then you're not actually aware and then you're not necessarily liable so that's helpful. Another avenue of which you become aware that people have mentioned is the Trusted Flagger Regime. So I think many organisations probably have bilateral arrangements with industry groups on a formal or informal basis.

[00:16:47.110] - Ciara Anderson, Senior Associate

This is now a legal standard to have these trusted flaggers, for example Europol, or organisations that identify child sex abuse material, and they'll be identified by the digital service coordinator. So that's another avenue of which you become actually aware. That's now been codified, if you will and then another area which we haven't spoken to yet is law enforcement orders. So that's another avenue by which you could become actually aware and again, organisations receive these orders. We see them all the time. They're in different languages, they're upside down, they're sideways, they don't have any information in them and organisations are really struggling to parse them and determine what kind of content they're referring to. So there's now, again a requirement that all member states have to set for what these law enforcement orders need to have in them. So, for example, they need to have who's the issuing authority? What is the legal basis for the order itself? What is an explanation of why the content is illegal with a reference to the particular law, and then just practical things like the territorial scope of the order, and it needs to be in the language of the provider or that they understand so they're not trying to translate.

[00:17:59.200] - Ciara Anderson, Senior Associate

So I think that's quite helpful for organisations who have ultimately had to develop these processes themselves. Two final points which I think have been mentioned is just, again, it's taken away the discretion in terms of the criminal offence. So many organisations would report to law enforcement if they're aware of information in relation to a criminal offence, just as good citizens. But now you're required

to do so if you're aware of information relating to a crime that has to do with the life or safety of an individual. So again, with the GDPR, you want certainty in relation to some of these situations and you mentioned the repeat offenders as well. Organisations may have something like that already in play, but this again is a requirement. If you have people who are frequently posting manifestly illegal content or abusing your complaint programme, and they're constantly making appeals or complaints that are completely unfounded, then you can, with a warning, essentially suspend them from the platform for a period of time. So the principles aren't changing, but there is a process there now that organisations can maybe take a little bit of the pressure off of having to develop these systems themselves

[00:19:13.970] - Ciara Anderson, Senior Associate

so I think it will be welcome.

[00:19:16.280] - Rob Corbett, Partner

Yeah, I think Helen mentioned earlier on that Technology Ireland are coming together and never mentioned as well the need for codes of conduct and so forth and those codes of conduct have to be informed by the reality of how all of these platforms are going to address all these new rules but at least there will be one set of rules, I suppose, and it'll apply on a Pan-European basis. So that will be helpful, maybe if we briefly flip it around for you, Rose. I mean, we've looked at it from the context of the platforms, but in terms of the people who are impacted by the allegations that they've posted content in breach of the new regime, it was mentioned earlier on in one of the questions, what are the implications for those people?

[00:19:53.630] - Rosemarie Blake, Senior Associate

Sure, Rob. So users are now going to have far greater insight into the decision making process which has led to their content being removed and the right to be provided with far more information surrounding the removal of content online. So if I'm a user of a platform, that platform's terms and conditions now have to tell me how decisions are made in the first instance, and if my content's been taken down, I've also got the right to be provided with a statement of reasons. It's gotten quite a lot of airtime today, and for good reason, because it's very prescriptive. So that's going to tell me why my content's been taken down as a consequence of one, another use of the platform, or two, an own volition investigation by the platform itself. So the statement of reasons includes a very prescriptive list of information, which includes the territorial scope of the content removal decision, the duration of the content removal, like how long it's going to last, whether AI was used to identify the content and disable it, and when the removal decision has been founded on illegality, the statement of reasons has to reference the law that's been contravened, and then for content that's incompatible with terms and conditions, the decision has to reference the contractual ground of those Ts and Cs that's been relied upon, and an explanation on why that content is incompatible with that provision.

[00:21:08.550] - Rosemarie Blake, Senior Associate

You also have to be provided with clear and user friendly information on the possibilities for redress that are available, the internal complaint handling mechanisms, like out of court dispute resolution and judicial address. So it is important to note, as a user, if you engage with the internal process or out of court dispute settlement, that doesn't in any way opt you out of your rights to go to court to obtain an order. I think within the recitals of the DSA, it says, none of this is without prejudice to a person's right to seek an effective remedy before the national courts. So then, looking at the complaints lodging process, a user can lodge a complaint appealing a removal decision for up to six months after the date when they're informed, and then the platform is required to handle the complaint in a timely, nondiscriminatory, diligent and non arbitrary manner, which requires three key things, reconsidering its decision regarding the information and potentially reversing it where there's a reasonable and legitimate reason to do so. Inform complainants without undue delay on their decision regarding the complaint and then around the AI you've got to ensure that decisions are made under the supervision of appropriately qualified staff and not solely on the basis of automated means.

[00:22:21.870] - Rosemarie Blake, Senior Associate

So for organisations mapping this impact, you may or may not currently offer an appeal option. So you'll need to identify resources to respond to those appeals, and then the appeal process, as I said, can't be wholly automated, so you've got to identify a human along that process, along that point to make sure there's appropriate oversight and quality control. There is an emphasis on costs here, which is meant to remove the cost burden for that user. So if the platform loses, there's a cost penalty, but no such corresponding burden on the user. So users are absolutely going to want to go and exhaust all those possible avenues to the fullest extent before resorting to court proceedings.

[00:23:02.090] - Rob Corbett, Partner

Yeah, I mean, that's interesting. So the availability of these avenues will be for all, at low cost, as opposed to currently, I think a lot of these redress mechanisms are only available for those who can put their hands in the pocket and effectively recruit lawyers to go off to the court for relief so there's going to be a lot of changes for content moderators under their DSA obligations, both to the recipients of the services and then to those who are seeking the removal of illegal content. Ciara, you mentioned that DSCs will be notified by law enforcement requests under the DSA, the platforms will be notified of these law enforcement and they'll have ongoing reporting and governance obligations. So maybe like on a business as usual basis, when we are live in this new regime, what are these reporting requirements going to look like?

[00:23:52.040] - Ciara Anderson, Senior Associate

Sure, I think some of them have been mentioned already, so I won't go back overground, but the transparency reporting is the biggest one and it's been mentioned before, but there's a lot of detail, there's a lot of data points, all the things that we've discussed here, but things like the timing it takes to respond, the decision, the type of content, how many decisions you overturn then. So if a decision comes in, you

decide to suspend or remove content, and then on the complaint handling process, you decide to reinstate. That needs to be reported because I think they're going to look at if basically a lot of decisions are being reversed. You have to also report on your repeat offenders and all this. So there's a lot of data points that are going to be collected and that's all publicly reported. There was another piece that I don't think has been mentioned, and I actually don't know if it's gotten a lot of airtime is there is a real time reporting obligation in relation to decisions and statement of reasons so that needs to go up to the commission and there's going to be a publicly available database of that information that was scrubbed of personal data, but that'll be a huge, vast amount of information in relation to how these decisions are being made

[00:25:02.140] - Ciara Anderson, Senior Associate

and again, there'll be some kind of engineering lift in order to collate that information and to send it to the commission. So those are the two main, I suppose, reporting obligations; the real time one, essentially and then the transparency report was obviously going to be a major operation. I won't go into detail on the risk assessment because it's already been mentioned. That isn't a reporting obligation. Ultimately, you undertake the risk assessment. VLOPs and VLOSEs, not everyone will have to do it, but that's not automatically reported but the commission, the DSC and the Digital Services Board can request information, and I'm sure they most definitely will, particularly because the digital services board also has their own report that they need to publish on a yearly basis in relation to what they're identifying as systemic risks and best practises and that information is going to come from what they're pulling from the risk assessments and also what they're seeing from some of this database of decisions as well so there's going to be a lot of information out there, which is a good thing.

[00:26:05.860] - Rob Corbett, Partner

Yeah. Look, hopefully there's been a lot of information in this room as well. So I'm conscious that we need to allow time for our final panel but maybe the key takeaway I'm taking from this one is that existing content moderation programmes are not perfect and are constant, kind of, iterative process of using technology to identify harmful, illegal content and taking it down. Get ready for a new regime that's a lot more prescriptive about what you must do and how you must do it, and also get ready for the reporting lift that goes with that, because you will be reporting to the very regulators who can fine you large amounts of money, as we'll hear in the last panel, in terms of how you account for your behaviour under the DSA and the OSMRA standards. So I might just draw this panel to a conclusion, because there will be time for questions before we break for lunch after the next panel. But for now, I'd just like to thank Olivia, Ciara and Rose for their thoughts on the content moderation.