GROK: CRIMINAL OFFENCES, THE CIVIL LAW AND THE ONLINE SAFETY ACT

This analysis by Prof Lorna Woods OBE looks at the use of Grok, X's AI tool, to create non-consensual, sexually explicit images of women and girls, the possible criminal offences that users might commit when using it for these purposes and how the Online Safety Act might apply.

Background

2026 has started with widespread outrage over the creation – at scale – of sexually explicit or "undressed" images of women and children by Grok, the AI tool integrated into X, in response to user requests. This is nothing new. Since at least May last year, tech press reporting has highlighted the potential problems. Inadequate safeguards allow the creation of images (usually of women) in a state of undress (transparent bikinis and the like), or changing the pose to a more sexualised one. Infamously, an image of Taylor Swift purported to show her in lingerie – an image that went viral. Civil society has also expressed concern, warning of the likely problems arising from "spicy" mode on "Grok Imagine". Some of the images appear to be of underage girls. Other earlier reports highlight the requests for child sexual abuse material (CSAM), suggesting that this may not be an isolated problem, as this overview of the issues by one of the journalists who's tracked the developments reaffirms.

What's changed in recent weeks is the introduction of the ability for any user to prompt Grok to edit images posted by other people on the platform, through an "edit image button": this removed friction from the process. One company (which detects manipulated images) at the end of 2025 estimated that there was a request for a "nudified" image per minute on Grok. The most recent response from X has been a post on the platform from Elon Musk, pointing out that users could be committing criminal offences; he did not acknowledge any responsibility on the part of X/Grok in this context.

Some countries have responded to this (Tech Policy Press is keeping a running list): in France, X was referred to the Media regulator, Arcom, in relation to contravention of the Digital Services Act and the existing investigation (relating to scams and foreign interference) into X is likely to be expanded. The European Commission has announced that it is looking into the allegations of CSAM. In India, the Ministry of Electronics and IT demanded that xAI explain — within 72 hours — what safeguards failed to prevent the spread of obscene and sexually explicit AI-generated material. The Malaysian authorities have also expressed concern. In the UK, Ofcom has been in "urgent contact" with X – though to what

end is unclear; the Home Office has pointed to proposals to ban nudification apps and the DSIT Secretary of State has made it clear that Ofcom has her backing to take full enforcement measures if necessary.

The following note provides an overview of possible criminal offences that users might commit when using Grok for the purposes described above, as well as considering how the Online Safety Act might apply. The discussion must be understood with the caveat that much may depend on the particular circumstances of an individual event and it is difficult to draw firm conclusions in the abstract.

Criminal Law

Creation of Deepfakes (Synthetic Images)

Children

It is illegal to create a "pseudo-photograph" – section 160 of the Criminal Justice Act 1988 makes the possession of an indecent photograph or "pseudo-photograph" of a child a criminal offence; section 1 Protection of Children Act 1978 (PCA) also creates possession and distribution offences in relation to photographs and "pseudo-photographs" of children. Section 7(7) PCA defines "pseudo-photograph" as "an image, whether made by computer-graphics or otherwise howsoever, which appears to be a photograph", which would in principle cover AI-generated or altered images such as those produced by Grok. Possessing and making have been understood broadly for the purposes of these offences, though there is still some uncertainty as to how they would be understood in this particular context. The Internet Watch Foundation is reported as saying, however, that in its view none of the images it has seen cross the criminal threshold – though it is not known which particular images it is referring to when it says that.

Adults

Section 138 of the Data (Use and Access) Act (DUAA) amended the Sexual Offences Act 2003 (SOA) to extend the offence of sharing an intimate image without consent (s 66D) by inserting new ss 66E-66H to provide for new offences relating to creating (66E), or requesting the creation of (66F), purported intimate images of an adult without consent or reasonable belief in consent; the other provisions provide a defence, definitions and time limits for prosecution.

The provision, and the sharing offence, cover synthetic media that replicates an adult person's face and/or body, most likely using artificial intelligence. Significantly, the image must be created intentionally. There is a question of whether this requirement is satisfied where there is more than one possible output. It could be argued that the examples included in the reporting on Grok's output could satisfy this test because the prompts seem specific as to the nature of the output. This might not always be the case. There seems to have been no suggestion that there were attempts to obtain consent, which could suggest that the requirement that the creation of the image was without consent or reasonable belief in there being consent. The Government's original proposal was to base the offence on the perpetrator's intent, which would have made the offence more difficult to show.

Finally, the image must be of an "intimate state" – as defined in relation to the existing intimate image offence in the SOA 2003 (s 66D(5)-(9)). Many of the images discussed in the reporting on Grok would seem to fall within the definition of intimate, as it includes situations where people are portrayed in

underwear or “transparent” bikinis. Presumably some images might not constitute an intimate state – for example, those where a bikini was not transparent – though even here, the images could fall within the regime if other content clues were sexually suggestive (eg pose). Section 138 DUAA has not yet been commenced, although the rules on sharing such images are. The sharing of images offence is in force and is listed as a priority offence in the Online Safety Act. (See below for detail on the regulatory requirements on platforms related to these offences.) It is not certain, however, that prompting an image which is then publicly displayed constitutes sharing such an image.

The Government is also introducing laws, via the Crime and Policing Bill currently before Parliament, to make it illegal to possess, create or distribute AI tools made or adapted for use for creating, or facilitating the creation of CSAM (or to provide guidance on how to use such technology to produce CSAM) (See cl 63 and cl 64). It has also announced that it will prohibit nudification apps, though precise details are not currently available. It is unclear whether the intention is to cover general purpose apps with insufficient safeguards or whether the focus is just on apps designed for the purpose of nudification (or even what nudification is – would it cover undressing people into bikinis?). Neither of these sets of provisions are in force yet.

There are other criminal offences which might be relevant, but which do not focus on the manipulation of images or necessarily on sexualised images. Note section 179 OSA covers sending messages the sender knows to be false, with intent to cause non-trivial psychological or physical harm to the likely audience, with no reasonable excuse for sending the message. Given that the requests for changes to photographs can show in the comments section (see s 179(2) OSA for definition of “likely audience”), it could be that amendments to images could fall within this offence. Ofcom has recognised the harm that intimate image abuse can cause in its Register of Risks. Clearly, instructing Grok to revise an image means that the user knows that the image is a fake; there would, however, be questions around whether the person requesting the image intended to cause non-trivial psychological harm. Alternatively, section 127(1) Communications Act applies when an offender sends, or causes to be sent, via a public communications network a communication that is either grossly offensive, or of an indecent, obscene, or menacing character. These offences may apply to some of the images, but their application is a little hard to predict. While the CPS Guidance notes the importance of freedom of expression, it also notes that “[o]nline activity is used to humiliate, control, and threaten victims”.

As ever with internet communications and the criminal law, there will be the issue of whether the UK authorities have jurisdiction. An offence must have a "substantial connection with this jurisdiction" for this to be the case. Substantial connection may be satisfied where either the suspect or the victim are here, though there might well be practical difficulties in taking a prosecution forward were a perpetrator to be overseas.

Civil Law

Individuals may also be able to rely on rights in the civil law. The alteration of a photograph will involve the processing of personal data, triggering the Data Protection Act 2018 rules. So, taking someone’s photographs from social media and using them to fabricate an image or video would violate UK GDPR principles – there is no consent here or a legitimate purpose. English law also protects against the misuse

of private information; it is aimed at redressing the violation of a person's intimacy and dignity, and may therefore be particularly relevant to the sexualised context. Moreover, a deepfake that falsely portrays someone in a manner that damages their reputation can give rise to a defamation claim under the Defamation Act 2013. The meaning attributed to the altered image will be central to any such claim's success. Copyright may be relevant when original works (photographs uploaded by the victim, where the victim is the copyright owner – eg a posed selfie) are altered by AI without authorisation. While there are "fair dealing" defences which might be appropriate for some deepfakes (including parody and satire) they seem unlikely to be relevant here. There is also likely to be an infringement of the moral rights of the victim as copyright creator. The common theme here is that the victim would be responsible for initiating and funding the action against the person who caused the image to be generated, which is a significant burden both financially and psychologically/emotionally.

Impact of the Online Safety Act Regulatory Regime

As noted, the duties in the regime relate to user-generated content which is "illegal content" and "content harmful to children". This is why the question of whether there is a criminal offence (and which criminal offence) remains important even in this regulatory regime. It follows that until an offence has come into force, it cannot trigger OSA duties – this is the case for the creation of deepfakes offence in DUAA.

User-to-user services, which includes X and Grok as part of X's service, have an obligation to have a system in place to swiftly take down illegal content (whether priority or not) once aware of it – Ofcom has left it to services to determine what "swiftly" means. Additionally, they are to mitigate and manage risks of harm identified in the service's risk assessment – this could cover harm from content associated with both priority and non-designated offences. The OSA specifies further obligations in relation to priority offences:-

- prevent individuals from encountering priority illegal content
- effectively mitigate and manage the risk of the service being used for the commission or facilitation of a priority offence and minimise the length of time for which any priority illegal content is present.

Many child sexual offences are priority offences, including s 1 PCA and s160 CJA mentioned above. Assuming that content produced by Grok in response to user prompts is user-generated content, some of that content could trigger the illegal content duties (as discussed above). Nonetheless, given the "safe harbour" provision (s 49 OSA), service providers need not do more than required by Ofcom's Codes, no matter what their risk assessments say. Currently, these focus on corporate governance of risk and content moderation, rather than technologies which cause or exacerbate the issue in the first place. **So, even if the content being produced triggers the duties (which some of it likely does), the Codes do not address the functionality that is central to these images being created.** This suggests that on its current iteration of Illegal Content Codes, Ofcom cannot hold X accountable for rolling out a function that has been used to generate illegal content and has allowed multiple users to commit criminal offences. Ofcom can only hold X to account for responding to user complaints about the content that has been created as a result. (There may also be failings around X's risk assessment process).

It is unclear whether these images would fall within the "content harmful to children" duties. While there is a general definition of harmful to children, the Act contains lists of primary priority content and priority content. Primary priority content includes pornography. Where the images are pornographic user-generated content, they should be age-gated. It is unclear whether this is the case for all the images. Section 236(1) OSA defines pornographic content as "content of such a nature that it is reasonable to assume that it was produced solely or principally for the purpose of sexual arousal". Ofcom distinguishes it from sexually suggestive or sexualised content, which is "material that implies or hints at sexual themes and/or intention, to evoke sexual thoughts or reactions, but that is not explicit and has not been produced solely or principally for the purpose of sexual arousal". Ofcom gives the example of someone playing a game in swimwear; it also suggests semi-nudity (other than people in underwear) will not constitute pornography (or at least not without other indicators, some of which may be present in at least some of the prompts/images).

If the duties are triggered, then it is unclear that Ofcom's Children's Code as yet addresses this issue directly, and expecting the services to take steps runs into the safe harbour problem noted above.

Where that leaves us

While we have noted that there are a number of civil law actions that could be relevant in this context, the OSA duties are not triggered by them but just the criminal law. This shows the limitations of fixating on the criminal law as describing all the issues that the law seeks to control offline; quite clearly, individuals' rights are in issue but the OSA regime does not offer any protection.

Ofcom has asked for an urgent response from X, and has said that it will act swiftly when it gets a response. It is unclear how long Ofcom will allow X to respond, before it takes any further steps – and how long Ofcom's own processes will take before it can reach a conclusion. Until Ofcom has reached that conclusion, X may continue to operate Grok as it has done so far – though presumably, were it to do so, this could be a factor that Ofcom could take into account in any fines or other action it decides upon.

Online Safety Act Network
January 2026