Running head: Judgements of Deepfake Media Production

Celebrity status, sex, and variation in psychopathy predicts judgements of and proclivity to generate and distribute deepfake pornography

Dean Fido[1]*, Jaya Rao[1], Craig A. Harper[2]

[1] University of Derby (UK)

[2] Nottingham Trent University (UK)

* Correspondence concerning this article should be addressed to Dr. Dean Fido, University of Derby, One Friar Gate Square, Agard Street, Derby, DE1 1DZ, UK. Tel: (01332) 597861. E-mail: deanfido.psych@gmail.com

Declaration of Competing Interests

The author(s) declare no potential competing interests with respect to the research, authorship, and/or publication of this article.

Funding

This work was supported by an Undergraduate Research Scholarship Scheme (URSS) fund (£200) from University of Derby.

Abstract

With the advent of means to generate and disseminate fake, sexualised images of others for the purposes of financial gain, harassment, or sexual gratification, there is a need to assess and understand the public's awareness and judgements of said behaviour. In two independently-sampled studies, we used moderation (Study 1; *n* = 290, 42% female) and linear mixed effects (Study 2; *n* = 364, 51% female) analyses to investigate whether judgements of deepfaking (measured across 12 self-report items) differed as a function of victim status (celebrity, non-celebrity), victim and participant demographics, and image use (sharing, own sexual gratification), whilst controlling for the potential covariates of psychopathy and beliefs about a just world. We consistently observed more lenient judgements of deepfake generation and dissemination for victims who were celebrities and male, and when images were created for self-sexual gratification rather than being shared. Moreover, lenient judgements, as well as proclivity to act were predicted by greater levels of psychopathy. We discuss our findings in the context of future research needing to better understand the general public's rationale for said disparity in judgements, as well as identifying and combating barriers to disclose victimisation. Open data and a preprint of this paper are available at https://osf.io/fp85q/?view_only=8006547d6a524f4fbb9dd55005c73319.

Keywords: *deepfake media production, non-consensual image-based offending, judgements, psychopathy*

Celebrity status, sex, and variation in psychopathy predicts judgements of and proclivity to generate and distribute deepfake pornography

The production and non-consensual dissemination of sexually explicit deepfake images, henceforth referred to as *deepfakes*, refers to the use of widely accessible artificial intelligence to dynamically transpose one image (or a series of similar images) onto a secondary source (e.g., a still or motion picture); giving the indistinguishable illusion that a target individual is engaging in sexual behaviour (Li et al., 2019; Rofer, 2016). Deepfaking sits alongside the non-consensual taking (upskirting), dissemination ('revenge pornography'[1]), and sending (cyber-flashing) of private sexual images under the umbrella term of image-based sexual abuse (McGlynn et al., 2017). Sexualised deepfakes are thought to originate from the now-removed Reddit thread 'r/deepfakes', before featuring on pornography websites such as PornHub.com (Cole, 2018). Even though such sites have previously commented that in line with their terms of service, they would remove any non-consensual content – including deepfakes (Cole, 2018) – at the time of writing, deepfakes featuring the likeness of Emma Watson, Gal Gadot, Taylor Swift, Meghan Markle, and Scarlett Johansson appear in search engine results; suggesting complacency in the policing of such material and/or the pervasiveness and frequency of their production and dissemination. Female celebrities are considered the primary victims of deepfake generation; with such images gaining internet notoriety through the sexual nature and celebrity status of the individuals they profess to portray (Citron & Chesney, 2019; Delfino, 2019). Although there exists no published research documenting rationale for engaging in such behaviour, Harper et al. (2021) hypothesise that perpetrators may be motivated by curiosity, addiction, and/or the pursuit of sexual gratification or financial gain. Within this manuscript, we present the first psychological investigation into judgements of, and proclivity towards, pornographic deepfake media production.

Protection for deepfake victims is important, as even though we have little understanding of the direct, personal consequences of deepfake victimisation specifically, it is well-established from the wider image-based sexual abuse literature that victims experience social consequences, such as embarrassment, breakdowns of relationships, and

[1] The authors recognise the negative discourse around the term 'revenge porn(ography)', in that it both places blame on the victim through the term *revenge* and also suggests that the acts depicted are always consensual through the term *pornography* (see Fido & Harper, 2020). However, this term has been used here as this is the more common and lay-term for such act.

damaged reputations (Bloom, 2014), professional consequences, such as employment termination due to potentially damaging an organization's reputation (Citron & Franks, 2014), and health consequences, such as facilitating depression, anxiety, and stress related to trust and self-image (Bates, 2017).

Although scholars have begun to empirically test public judgements towards image-based sexual offending (e.g., Fido et al., 2021), there is a noticeable gap in the literature regarding judgements of deepfake offending, specifically, as well as how said judgements might be influenced by psychological factors. Across two samples based within the UK, Fido et al. (2021) used moderated moderation analyses (controlling for empathy-related personality traits) to evidence a trend for leniency in judgements of revenge pornography to be predicted by self-reported intrasexual competition. Although a positive first step, one factor that the authors did not examine was whether judgements differed as a function of the celebrity status of the victim – relevant given the prevalence of high-profile victims such as actors Jenifer Lawrence, Kate Upton, and Kirsten Dunst (to name just a few). Celebrities are increasingly becoming the targets of trolls and online abuse (Dooley et al., 2009; Garde-Hansen & Gorton, 2013), including being targets of deepfaking. But even though we know that celebrities *feel* the impact of antisocial online behaviour (Ouvrein et al., under submission), investigations into perceptions of different victims is lacking (Scott et al., 2020).

We know that some of the hostility towards celebrities might emerge from the use of their platform to voice controversial and/or outspoken views or gossip (Muntean & Petersen, 2009). Recently, Scott et al. (2020) suggested that such messages, measured via tweets, might even act to attenuate the perceived impact of abuse from others, relative to if a celebrity has received abuse after tweeting positive content – potentially as a function of responders attributing some of the blame for the abuse to the victims (Scott et al., 2018). Moreover, responders may perceive a reduced impact of internet-mediated abuse on celebrities due to beliefs that such actions are part and parcel of being famous (Ouvrein et al., 2017; Ouvrein et al., 2018) or that they are simply desensitised to such impacts due to shifts in social norms within online environments (Pabian et al., 2016; Pornari & Wood, 2010). Indeed, responders report feeling *safer* when negatively commenting about celebrities, relative to non-celebrities (Feasey, 2008). Together, this acts to proliferate the conception that celebrities hold a social position that is distant from the general population (Henrich & Gil-White, 2001) - potentially decreasing our empathy for them, relative to non-celebrity victims (Peng et al., 2015).

Such findings are of particular interest to image-based sexual abuse, whereby even though victims are not held fully accountable for the actions of others (Eikren & Ingram-

Waters, 2016), their past actions and behaviours are scrutinized. Evidence towards this stemming from the revenge pornography literature includes perceptions that the initial self-taking of sexual images was 'reckless', 'careless', and 'naive' (Henry & Powell, 2015), that individuals should be aware of the risks posed to them (Henry et al., 2017), and that the management of such images (including their deletion) should be dealt with by the image-taker after the breakdown of a relationship (Gavin & Scott, 2019). Despite overlaps between both the platforms and mechanisms by which offending takes place, as well as the invasive and pervasive impact victimisation has on its victims (McGlynn et al., 2017), it would be complacent not to explore public judgements of deepfakes in a direct way. Moreover, and along the lines of both Pina et al. (2017) and Fido et al. (2021), such judgements may be impacted by individual differences between observers, such as those pertaining to callous and unemotional responses, and so two potential covariates are outlined below.

Psychopathy

Within the general population, psychopathy is thought to manifest on a continuum, with high scorers characterised by a constellation of shallow emotion processing, inappropriate affective, and a reduced capacity to experience empathy (Viding & McCory, 2019). Such individuals are also considered to be at greater risk of engaging in aggressive and antisocial behaviour (Blais et al., 2014; Marsh & Cardinale, 2012) as well as reporting higher levels of sexual harassment proclivity (Zeigler-Hill et al., 2016). In an online arena, psychopathy is positively associated with engagement in a higher frequency of trolling behaviours (Buckels et al., 2014), endorsement of unprovoked celebrity-focused aggression (Scott et al., 2020), and an increased use of profane language (Sumner et al., 2012) and cyberaggression (Pabian et al., 2015). Moreover, it predicts proclivity to both seek revenge following infidelity within a relationship (Brewer et al., 2015), and to commit revenge pornography offences (Pina et al., 2017) - potentially as a function of pleasure gleaned from inflicting emotional distress on others (Kircaburun et al., 2018; Sest & March, 2017). Thus, psychopathic personality traits provide a potential individual difference worth controlling for when assessing judgements of deepfakes.

Belief in Just World

As described, there is a pervasive position of image-based sexual offending, and to some extent sexual offending more generally to place blame for victimisation on the victims themselves (Henry et al., 2017; Henry & Powell, 2015). Victim blaming in this capacity can

occur as a function of believing that the world is a fair place where people deserve what they get, and get what they deserve (Lerner & Simmons, 1966), and underpins what are commonly referred to as 'rape myths'; a set of beliefs about how victims of rape might have contributed to their own victimisation (Vonderhaar & Carmody, 2015). As such, it makes sense to assess whether one's disposition to believe in a just world (Wenzel et al., 2017) would impact deepfake-related judgements, especially in the context where some victims, for example celebrities, are thought to contribute to being victims of abuse across online settings (Muntean & Petersen, 2009).

Overview of Studies

To our knowledge, this is the first empirical investigation into the psychological predictors of judgements of the production and dissemination of sexually explicit deepfake images. Here, we present two explorative studies to begin to fill this gap in the literature. In the first study, we used a cross-sectional and experimental design with a moderation analysis to test whether judgements of a deepfake case would be more or less lenient if the victim was a celebrity (relative to a barista) or was male (relative to female). The covariates of psychopathy and belief in a just world were controlled for. In the second study (ran parallel to the first), we conducted an extension of this work, switching from a between-subjects to a within-subjects design and experimentally manipulating whether deepfake images were produced for personal sexual gratification, or whether they were also later disseminated.

Study 1

Methods

Participants

For both studies, determinants of sample sizes, data exclusions (if applicable), and experimental manipulations are reported. We conducted an a priori power analysis using G*Power (version 3.1.9.2). Assuming an anticipated small-to-medium effect size (.09; Cohen, 1988) – to provide practical importance – and a standard alpha level of .05, a minimum of 195 participants were required to have 95% power in our planned analyses in each of the two studies using the *F* tests family (accounting for predictor and moderator variables, as well as covariates). To mitigate against incidents of missing data and participant withdrawal, we sought to recruit upwards of 250 participants. After removing cases where more than 5% of the data were missing (n = 4), a total of 290 UK-based participants (M_{age} = 34.83 years, SD = 12.51; 42% female) completed an online questionnaire, which was

advertised through the crowdsourcing website *Prolific*. Within this sample, men (M = 36.12 years, SD = 12.93) were significantly older than women (M = 33.05 years, SD = 11.73), $t(288) = 2.07$, $p = .039$, $d = 0.25$. This approach boasts comparable data to that obtained through lab and face-to-face means (Peer et al., 2017). Participants indicated their consent at the start and end of the online survey, and were reimbursed with £0.75 for their participation.

Materials

Demographics. Participants were asked to report their age and sex.

Self-Report Psychopathy Scale - Short Form (SRP4; Paulhus et al., 2014). The SRP4 comprises 29 items that measure psychopathic personality in both forensic, and non-forensic populations using a 5-point self-report scale (anchored from *"Disagree Strongly"* to *"Agree Strongly"*). High scores indicated greater levels of psychopathy one four different components: 'interpersonal' (e.g., "I purposely flatter people to get them on my side"; Cronbach's α = 0.79[2]; 7 items); 'affective' (e.g., "People sometimes say that I'm cold hearted"; Cronbach's α = 0.72; 7 items); 'lifestyle' (e.g., "I rarely follow the rules"; Cronbach's α = 0.77; 7 items); and 'antisociality' (e.g., "I have tricked someone into giving me money"; Cronbach's α = 0.72; 8 items).

Judgements of Deepfakes. Participants were asked to read one of four randomly presented vignettes outlining a deepfake incident involving an individual generating and disseminating a fake sexualised image of another after being unable to engage then in a physical relationship. Afterwards, they were asked to answer 12 judgement items adapted from Krahé et al.'s (2007) work on victim blaming in cases of rape; distributed across constructs of *victim blame* (e.g., "How much do you think [victim's name] is to blame for the incident?"; Cronbach's α = 0.86; 3 items), *criminality of the behaviour* (e.g., "Do you think police intervention is necessary for the resolution of the situation?"; Cronbach's α = 0.79; 3 items) and perceived *victim harm* (e.g., "Do you think [perpetrator's name]'s actions will create fear in [victim's name]?; Cronbach's α = 0.91; 3 items). Responses were made on a seven-point scale, anchored from "*Not at all/Very unlikely"* to "*Definitely/Very likely"*. Average scores were computed for each sub-scale, with high scores indicating more victim

[2] This Cronbach's alpha, as well as other presented in this manuscript are derived from the current datasets.

blame, judgements of the criminal nature of deepfake pornography offending, and perceived victim harm.

The four vignettes differed as a function of sex (male vs. female) and social position (celebrity vs. non-celebrity) of the victim. For non-celebrity victims, they were written as working a job as a barista. The character names Taylor and Ashley were deliberately used due to being unisex, so that the vignettes could be better standardised to reduce extraneous variance. The exact wording of the vignettes was as follows:

Taylor is an adult (male/female), who over the past month has developed a strong attraction to Ashley, a (male/female) (barista at their local coffee shop/celebrity who is famous for acting and singing). Lately, Taylor has begun to sexually fantasise about Ashley. Unable to strike up a physical relationship with Ashley, and to satisfy their sexual desires, Taylor generated fake pornographic videos of Ashley by superimposing (his/her) face onto the body of an adult porn star. Taylor then proceeded to share those videos with their friends.

Belief in Just World (BJW; Lipkusa et al., 1996). BJW was measured using Lipkusa et al.'s (1996) 8-item index of the same name (e.g., "I feel that the world treats people fairly") across a six-point scale ranging from *"Strongly Disagree" to "Strongly Agree"* (Cronbach's $\alpha = 0.89$). High scores indicated an increased belief that the world is a fair and just place.

Proclivity to offend and knowledge of offending. Participants' proclivities to [1] create, and [2] disseminate deepfake pornography were assessed using two single-item percentage-based scale, ranging from 0% (*Extremely unlikely*) to 100% (*Extremely likely*). These scales were prefixed with the following context, and participants were told that in this context, were would be *"absolutely no chance"* of getting caught:

You have access to pictures of somebody who you find very attractive in your social circle. They are not interested in entering into a sexual relationship with you, but you learn about software that is freely available which would allow you to superimpose the head of that person onto the body of an adult porn star; generating a realistic, but fake, pornographic movie featuring that person.

Finally, in an open text box, participants were asked to write down the name of the subject matter which had been discussed throughout the vignettes and the proclivity measures to gauge their knowledge of the behaviour. Of importance, the term 'deepfake' (or similar) was not used throughout the survey, except for the debrief document.

Procedure

After clicking on the study link, participants initially entered their demographic information, prior to completing the SRP4, BJW, and deepfake judgement vignette. These measures were randomly presented to reduce potential order effects. On average, the study took around 12 minutes to complete. This procedure was approved by an institutional review committee prior to data collection.

Results

In the analysis for each dependent variable, we ran a moderated moderation using the GAMLj module in jamovi (Gallucci, 2019). In this analysis, we used 'victim celebrity status' as our focal predictor of each dependent variable, with his relationship being moderated by 'victim sex', and further by 'participant sex'. All regression coefficients for moderated moderation models reported in this paper are unstandardized in line with Hayes (2018) and were bootstrapped using 5000 re-samples. Confidence intervals were not bias corrected. A schematic of this model is presented in Figure 1, and all outputsare available via the Open Science Framework project page (https://osf.io/fp85q/?view_only=8006547d6a524f4fbb9dd55005c73319).

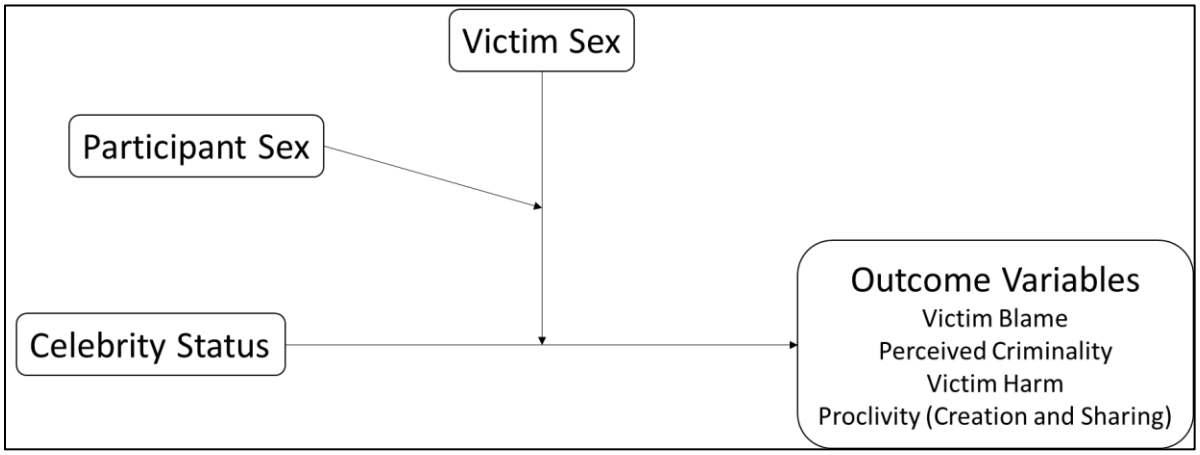

Figure 1. Planned analytic model

We first ran correlational analyses to ensure that our key variables were associated with one another. The correlation matrix is presented in Table 1. Examining this matrix, we see expected correlations between our outcome variables (victim blame, perceived criminality, perceived victim harm, and deepfake pornography proclivity) and psychopathic personality traits, with higher psychopathy being associated with higher levels of victim blame, lower perceptions of criminality and victim harm, and a greater proclivity towards creating and sharing deepfake pornography. Belief in a just world was associated with higher levels of victim blame, and a proclivity towards sharing deepfake pornography. All outcomes were moderately associated with each other (victim blame being positively associated with proclivity outcomes but negatively associated with perceived criminality and victim harm, which were positively correlated with each other).

Table 1. Pearson correlations between study variables

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1. Participant Sex	-										
2. Belief in a Just World	-0.06	-									
3. SRP Interpersonal	-0.23***	-0.05	-								
4. SRP Affective	-0.33***	-0.01	0.72***	-							
5. SRP Lifestyle	-0.28***	-0.01	0.67***	0.63***	-						
6. SRP Antisocial	-0.20***	0.10	0.47***	0.50***	0.55***	-					
7. Victim Blame	-0.05	0.18**	0.03	0.10	0.10	0.23***	-				
8. Perceived Criminality	0.26***	-0.05	-0.20***	-0.22***	-0.16**	-0.11	-0.30***	-			
9. Victim Harm	0.21***	-0.09	-0.16**	-0.21***	-0.14*	-0.13*	-0.26***	0.68***	-		
10. Proclivity - Creation	-0.32***	0.04	0.45***	0.34***	0.37***	0.27***	0.14*	-0.23***	-0.28***	-	
11. Proclivity - Sharing	-0.11	0.13*	0.33***	0.26***	0.26***	0.42***	0.16**	-0.07	-0.12*	0.41***	-

* $p < .05$ ** $p < .01$ *** $p < .001$

Victim Blame

The model for victim blame explained a statistically significant proportion of the variance in this outcome variable, $R^2 = .125$, $F(12, 276) = 3.29$, $p < .001$. Model coefficients are presented in Table 2.

Table 2. Model coefficients for victim blaming

Names	*b*	*SE*	95% CI Lower	95% CI Upper	β	*t*	*p*
(Intercept)	1.59	0.07	1.46	1.72		23.92	< .001
Celebrity Status	0.04	0.14	-0.22	0.31	0.04	0.33	.743
Victim Sex	**-0.29**	**0.13**	**-0.55**	**-0.02**	**-0.25**	**-2.15**	**.033**
Participant Sex	0.01	0.14	-0.27	0.29	0.01	0.05	.957
Belief in a Just World	**0.03**	**0.01**	**0.01**	**0.05**	**0.15**	**2.65**	**.009**
SRP Interpersonal	-0.03	0.02	-0.07	0.01	-0.12	-1.31	.192
SRP Affective	0.01	0.02	-0.04	0.06	0.04	0.44	.660
SRP Lifestyle	0.00	0.02	-0.04	0.04	0.01	0.12	.904
SRP Antisocial	**0.09**	**0.02**	**0.04**	**0.14**	**0.26**	**3.61**	**< .001**
Celebrity Status × Victim Sex	-0.33	0.27	-0.86	0.21	-0.28	-1.21	.228
Celebrity Status × Participant Sex	**-0.55**	**0.27**	**-1.08**	**-0.03**	**-0.48**	**-2.07**	**.039**
Victim Sex × Participant Sex	0.30	0.27	-0.22	0.83	0.26	1.14	.256
Celebrity Status × Victim Sex × Participant Sex	-0.17	0.54	-1.22	0.89	-0.14	-0.31	.758

Within this model, victims being female (b = -0.33, p = .033) predicted less victim blame, and belief in a just world (b = 0.03, p = .009) and the antisociality component of psychopathy (b = 0.09, p < .001) predicted more victim blaming. The 'Celebrity Status × Participant Sex' interaction was also significant (b = -0.55, p = .039). Examining the plot of this interaction (Figure 2), we see that men blamed celebrities slightly more than non-celebrities for becoming victims of deepfake pornography production (b = 0.32, p = .067), while women demonstrated the opposite trend (b = -0.23, p = .256), though neither gradient was statistically significant. In both cases, however, the differences appear small. No other variables were significantly associated with victim blaming.

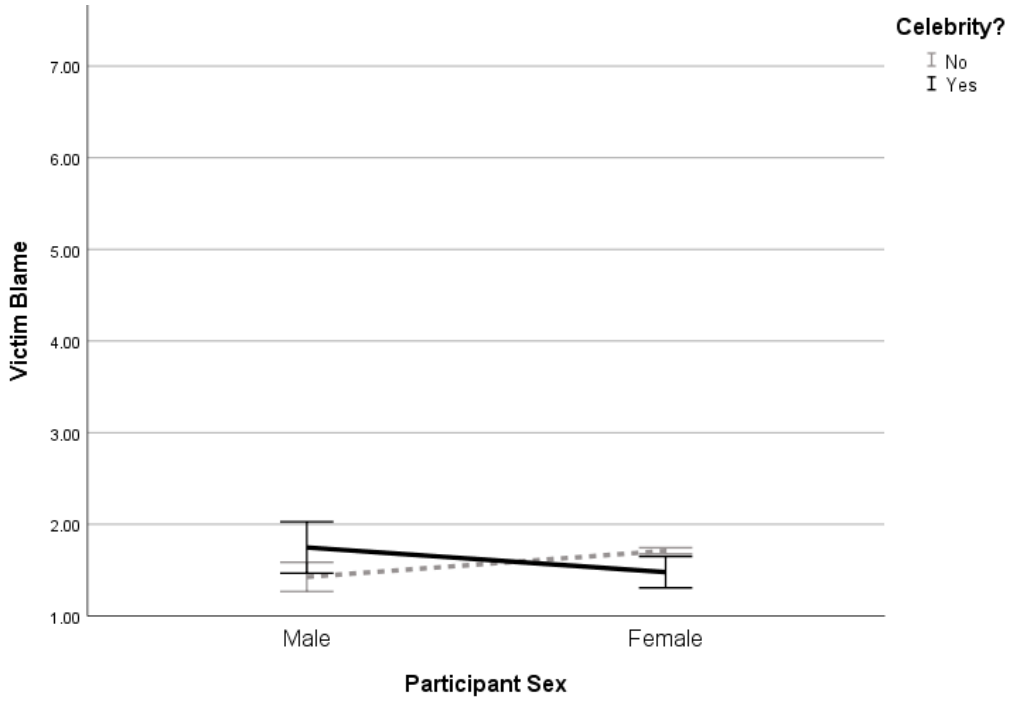

Figure 2. Celebrity Status * Participant Sex interaction, predicting victim blame

Criminality Judgements

The model for judgements of the criminality of deepfake pornography production explained a statistically significant proportion of the variance in this outcome, $R^2 = .146$, $F(12, 276) = 3.92$, $p < .001$. Model coefficients are presented in Table 3.

Table 3. Model coefficients for criminality judgements

			95% CI				
Names	***b***	***SE***	**Lower**	**Upper**	**β**	***t***	***p***
(Intercept)	5.50	0.08	5.35	5.66		68.86	< .001
Celebrity Status	-0.23	0.16	-0.55	0.09	-0.16	-1.43	.154
Victim Sex	**0.48**	**0.16**	**0.16**	**0.79**	**0.34**	**2.95**	**.003**
Participant Sex	**0.60**	**0.17**	**0.27**	**0.94**	**0.43**	**3.55**	**< .001**
Belief in a Just World	-0.01	0.01	-0.03	0.02	-0.04	-0.64	.526
SRP Interpersonal	-0.03	0.03	-0.08	0.02	-0.11	-1.28	.203
SRP Affective	-0.01	0.03	-0.07	0.04	-0.05	-0.52	.606
SRP Lifestyle	0.01	0.02	-0.04	0.06	0.04	0.48	.634
SRP Antisocial	-0.01	0.03	-0.07	0.05	-0.02	-0.30	.767
Celebrity Status × Victim Sex	-0.01	0.32	-0.65	0.62	-0.01	-0.04	.965
Celebrity Status × Participant Sex	**0.64**	**0.32**	**0.01**	**1.27**	**0.45**	**2.00**	**.046**
Victim Sex × Participant Sex	-0.35	0.32	-0.98	0.28	-0.25	-1.09	.278
Celebrity Status × Victim Sex × Participant Sex	-0.60	0.64	-1.87	0.66	-0.43	-0.94	.351

Within this model, deepfake pornography depicting female victims was associated with greater levels of criminality than cases with male victims (b = 0.48, p = .002), with women also judging deepfake pornography production as more criminal than did men in our sample (b = 0.60, p = .001). The 'Celebrity Status × Participant Sex' interaction was also statistically significant (b = 0.64, p = .046). Examining this interaction plot (Figure 3), we see that men judged deepfake pornography cases involving celebrity victims as being less criminal than cases involving non-celebrity victims (b = -0.55, p = .009), whereas women did not differentiate between these victim groups when judging criminality (b = 0.09, p = .717). No other variables were significantly associated with criminality judgements.

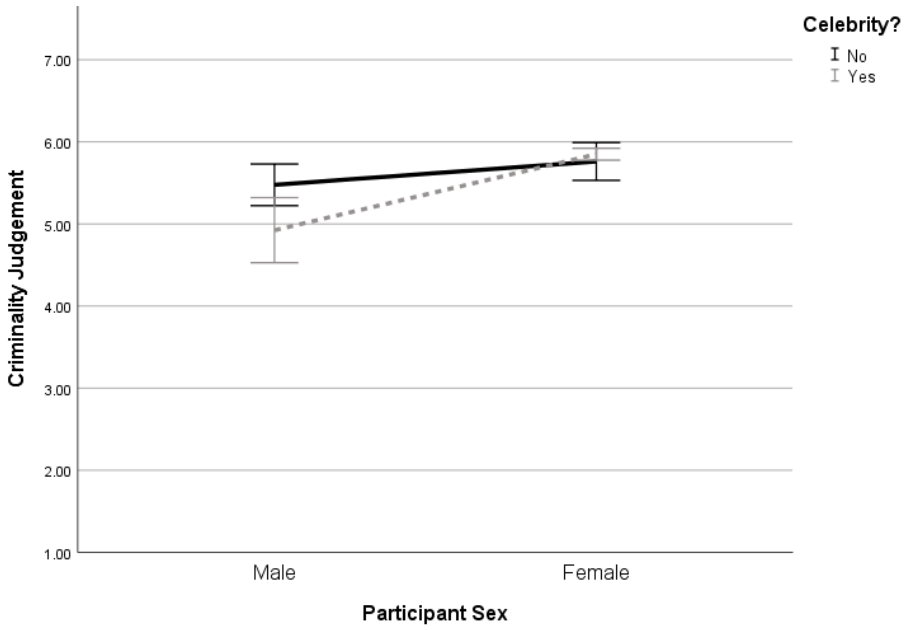

Figure 3. Celebrity Status × Participant Sex interaction, predicting criminality judgements

Victim Harm Perceptions

The model for perceptions of victim harm explained a statistically significant proportion of the variance in variable, R^2 = .191, F(12, 276) = 5.45, p < .001. Model coefficients are presented in Table 4.

Table 4. Model coefficients for victim harm perceptions

Names	*b*	*SE*	95% CI		β	*t*	*p*
			Lower	Upper			
(Intercept)	5.72	0.07	5.58	5.86		81.60	< .001
Celebrity Status	**-0.49**	**0.14**	**-0.77**	**-0.20**	**-0.38**	**-3.41**	**< .001**
Victim Sex	**0.51**	**0.14**	**0.23**	**0.79**	**0.40**	**3.59**	**< .001**
Participant Sex	**0.45**	**0.15**	**0.15**	**0.74**	**0.35**	**2.99**	**.003**
Belief in a Just World	-0.02	0.01	-0.04	0.00	-0.09	-1.55	.122
SRP Interpersonal	-0.02	0.02	-0.06	0.03	-0.07	-0.83	.405
SRP Affective	-0.02	0.02	-0.07	0.03	-0.06	-0.70	.488
SRP Lifestyle	0.02	0.02	-0.03	0.06	0.07	0.82	.414
SRP Antisocial	-0.03	0.03	-0.08	0.02	-0.07	-1.08	.279
Celebrity Status × Victim Sex	0.45	0.28	-0.11	1.01	0.35	1.58	.116
Celebrity Status × Participant Sex	**0.66**	**0.28**	**0.11**	**1.22**	**0.52**	**2.36**	**.019**
Victim Sex × Participant Sex	-0.30	0.28	-0.85	0.25	-0.24	-1.08	.282
Celebrity Status × Victim Sex × Participant Sex	-0.67	0.56	-1.78	0.44	-0.53	-1.19	.235

Within this model, greater levels of victim harm were attributed when cases involved non celebrity victims (b = -0.49, p < .001) and female victims (b = 0.51, p < .001), as well as by female participants (b = 0.45, p = .003). The 'Celebrity Status × Participant Sex' interaction was also significant (b = 0.66, p = .019). Examining the plot of this interaction (Figure 4), it appears that female participants did not alter their perceptions of victim harm as a function of the victim's celebrity status (b = -0.15, p = .476). However, men attributed lower levels of harm when victims were celebrities (b = -0.82, p < .001). No other variables were significantly associated with perceptions of victim harm.

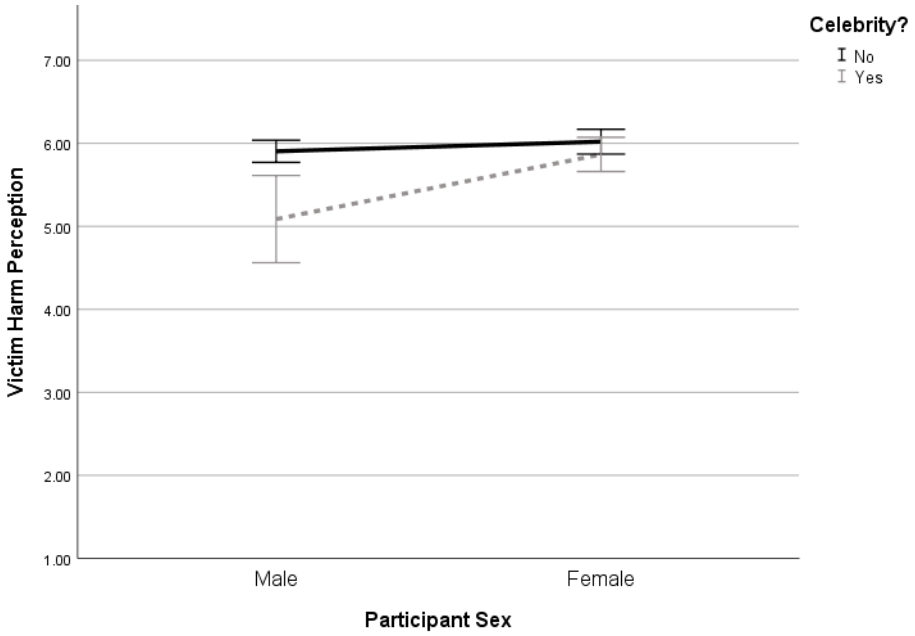

Figure 4. Celebrity Status × Participant Sex interaction, predicting perceived victim harm

Deepfake Pornography Creation Proclivity

The model predicting a proclivity towards creating deepfake pornography explained a statistically significant proportion of the variance in this outcome, $R^2 = .282$, $F(12, 274) = 8.97$, $p < .001$. Model coefficients are presented in Table 5.

Table 5. Model coefficients for deepfake pornography creation

			95% CI				
Names	***b***	***SE***	**Lower**	**Upper**	**β**	***t***	***p***
(Intercept)	12.01	1.28	9.49	14.52		9.40	< .001
Celebrity Status	5.09	2.60	-0.02	10.20	0.21	1.96	.051
Victim Sex	0.36	2.58	-4.72	5.44	0.01	0.14	.889
Participant Sex	**-11.81**	**2.72**	**-17.18**	**-6.45**	**-0.48**	**-4.34**	**< .001**
Belief in a Just World	0.18	0.19	-0.19	0.55	0.05	0.96	.340
SRP Interpersonal	**1.99**	**0.42**	**1.17**	**2.80**	**0.39**	**4.78**	**< .001**
SRP Affective	-0.62	0.44	-1.49	0.26	-0.11	-1.39	.166
SRP Lifestyle	0.36	0.40	-0.42	1.14	0.07	0.92	.361
SRP Antisocial	0.39	0.47	-0.55	1.32	0.05	0.81	.416
Celebrity Status × Victim Sex	-0.46	5.19	-10.67	9.76	-0.02	-0.09	.930
Celebrity Status × Participant Sex	-5.93	5.13	-16.02	4.17	-0.24	-1.16	.249
Victim Sex × Participant Sex	-3.47	5.11	-13.52	6.59	-0.14	-0.68	.498
Celebrity Status × Victim Sex × Participant Sex	-15.71	10.28	-35.95	4.52	-0.64	-1.53	.127

Within this model, men expressed a higher level of proclivity than women (b = -11.81, p < .001). Similarly, higher scores on the interpersonal component of psychopathy predicted a greater level of proclivity (b = 1.99, p < .001). No other main effects or interactions were statistically significant in predicting a proclivity for creating deepfake pornography.

Deepfake Pornography Sharing Proclivity

The model for deepfake pornography sharing explained a statistically significant proportion of the variance in this outcome variable, R^2 = .220, F(12, 275) = 6.45, p < .001. Model coefficients are presented in Table 6.

Table 6. Model coefficients for deepfake pornography sharing

			95% CI				
Names	***b***	***SE***	**Lower**	**Upper**	**β**	***t***	***p***
(Intercept)	2.18	0.49	1.22	3.14		4.46	< .001
Celebrity Status	0.57	0.99	-1.38	2.52	0.06	0.57	.568
Victim Sex	-0.70	0.99	-2.64	1.24	-0.08	-0.71	.478
Participant Sex	-0.38	1.04	-2.43	1.67	-0.04	-0.37	.712
Belief in a Just World	**0.14**	**0.07**	**0.00**	**0.29**	**0.11**	**1.99**	**.047**
SRP Interpersonal	**0.47**	**0.16**	**0.16**	**0.79**	**0.25**	**2.98**	**.003**
SRP Affective	-0.16	0.17	-0.49	0.18	-0.08	-0.94	.350
SRP Lifestyle	-0.13	0.15	-0.42	0.17	-0.07	-0.83	.405
SRP Antisocial	**0.99**	**0.18**	**0.64**	**1.35**	**0.37**	**5.48**	**< .001**
Celebrity Status × Victim Sex	-1.27	1.98	-5.17	2.63	-0.14	-0.64	.522
Celebrity Status × Participant Sex	-0.23	1.96	-4.09	3.62	-0.03	-0.12	.906
Victim Sex × Participant Sex	0.00	1.95	-3.84	3.84	-0.00	0.00	.999
Celebrity Status × Victim Sex × Participant Sex	-4.89	3.93	-12.62	2.84	-0.54	-1.25	.214

Within this model, none of our experimental manipulations predicted any variation in deepfake pornography sharing. However, believing in a just world (b = 0.14, p = .047), and both the interpersonal (b = 0.47, p = .003) and antisociality components of psychopathy (b = 0.99, p < .001) predicted a greater self-reported proclivity towards sharing deepfake pornography. No other variables, and none of the interactions, were statistically significant.

Labelling 'Deepfake' Media

We coded participants' open-ended responses when asked for the name of the type of behaviour depicted in the vignette. We coded the response as 'correct' if it included the word

'deepfake'. Of the 290 participants, 19 correctly identified the name of this behaviour. This corresponds to a 6.6% accuracy rate.

Study 2

Building on Study 1, we sought to replicate our initial findings in a second sample and explore whether such associations and judgements were further manipulated by whether deepfake images were solely generated for personal use or were subsequently disseminated. Although both actions carry potential moral implications, the former – especially when the victim remains unaware - is unlikely to bring about personal, professional, and social implications such as damaged reputations, embarrassment, and health consequences like depression, anxiety, and poor levels of trust (Bates, 2017; Bloom, 2014; Citron & Franks, 2014). The distribution of deepfaked sexual images, regardless of them being fake, would seemingly map on well to the growing literature pertaining to the non-consensual distribution of sexual imagery (see Henry & Powell, 2018 for a review) and may therefore add a subsequent route of victimisation in the absence of an initial sexual image or video being consensually sent. In addition, sharing such materials is a criminal offense in some jurisdictions (e.g., Abusive Behaviour and Sexual Harm Act (Scotland) (2016)). As such, there are both theoretically interesting and legally important considerations when considering judgments of and proclivities for personal use and shared deepfake content.

Methods

Participants

A total of 364 UK participants (M_{age} = 27.69 years, *SD* = 6.60; 51% female) completed an online questionnaire. Within this sample, men (*M* = 27.50 years, *SD* = 7.08) did not differ in age from women (*M* = 27.90 years, *SD* = 6.07), $t(361) = 0.60$, $p = .547$, $d = 0.06$. Of these participants, 62.6% were in a relationship. Owing to the small proportion of the sample that stated they were exclusively or predominantly homosexual (*n* = 26; 7.1%) we were unable to compare sexual orientation groups. Instead, we grouped participants according to the sex of their primary sexual attraction. We had an exact 50:50 division of the sample into 'androphilic' (exclusively or predominantly attracted to men) and 'gynephilic' (exclusively or predominantly attracted to women) clusters.

These participants were recruited via *Prolific*, with a blacklist being created to prevent those participating in Study 1 from accessing the link for Study 2. As such, the two studies

have completely independent samples. Participants were reimbursed with £0.75 for their participation.

Materials

Materials were the same as those outlined in Study 1, save for different vignettes being used as the experimental manipulation (see below). As in Study 1, the BJW scale ($\alpha = .86$), and the four domains of the SRP4 (interpersonal $\alpha = .80$; affective $\alpha = .72$; lifestyle $\alpha = .77$; antisocial $\alpha = .67$) demonstrated acceptable-to-strong levels of internal consistency.

We also asked participants to report their relationship status (in a relationship vs. not) and sexual orientation (exclusively/predominantly heterosexual vs. exclusively/predominantly homosexual) in Study 2. We did this to control for relationship status, and to allocate participants to vignette scenarios congruent with their sexuality.

Vignettes. We wrote 12 scenarios depicting the production of deepfake pornography for the purposes of this study. These were divided into six conditions, depending on the celebrity status of the victim (celebrity vs. friend vs. stranger) and the hypothetical use of the deepfake pornography (personal sexual gratification vs. sharing with others). Each of these six conditions had an androphilic and a gynephilic version, leading to the 12 vignettes. Below are the three gynephilic scenarios depicting personal sexual gratification as the motivation for deepfake pornography production. All scenarios are available via our Open Science Framework project page at https://osf.io/fp85q/?view_only=8006547d6a524f4fbb9dd55005c73319.

Celebrity

Mila Kunis is a leading film star and one of the most sought after actresses in the current film industry. You are that much of a fan of hers that you are always at the cinema to see the first showing of her new films, and regularly watch the movies and TV shows that she has starred in on Netflix. Mila is your ideal woman. You think she is the hottest woman in the world, and would love to date her. However, given her status you realise would never be possible. Due to this, you decide to edit Mila's face onto a pornographic video to feel like you can get close and intimate with her. Using new technology you decide to morph her into some of your favourite porn scenes, and use these for your own personal sexual gratification.

Friend

You and your best friend of nearly 10 years, Sarah, are keen to get away for a summer break with a group of friends as a school reunion, but arguments occur due to not being able to find a suitable destination for everyone. After months of inaction, you and Sarah decide to go away by yourselves to save the hassle. You choose to go to Santorini in Greece for a week in the middle of July. You and Sarah have never had a romantic relationship with one another, but did share a kiss when drunk around a year ago. You both agreed that this was just a funny accident, but you have always secretly had feelings for him. You found some software recently that allowed you to take pictures of Sarah's face from social media and morph them onto some of your favourite porn scenes. You have been using the edited porn videos for your own personal sexual gratification ever since.

Stranger

One evening while scrolling through social media sites, you come across a profile of a woman that you don't know called Sophie. She looks to be about 27 years old, and posts pictures of herself regularly. Instantly you find her sexually attractive and start to follow her account. You don't have any mutual friends, and while she doesn't have that many followers, it is unlikely she notices your account. As time passes, you find yourself going back to Sophie's account to see her updates and look at her pictures. One day you see a news article about photo editing programmes that allow users to morph images together. You download the program and find it works for videos, too. When you find this out, you decide to take a picture of Sophie and morph her face into one of your favourite porn scenes, and use this for your own sexual gratification regularly over the coming weeks.

After each presented vignette, participants completed the same victim blame, criminality, and victim harm perception scales as in Study 1, and the single-item proclivity scale ranging from 0-100. Average scores and internal consistency coefficients for each condition, by sexuality group, are presented in Table 7 (see Results section, below).

Procedure

The procedure for this study mirror that of Study 1. The only deviation was a switch from a between-subjects design to a within-subjects design. In practical terms, this means that

all participants responded to all six vignettes that were applicable to their sexuality (as determined by their self-declared sex, and the response to the sexual orientation demographic question), rather than just one. This allowed us to use each participant as their own control when comparing responses to each scenario. On average, the study took around 16 minutes to complete. This procedure was approved by an institutional ethics review committee prior to data collection.

Results

We used the *GAMLj* module in jamovi (Gallucci, 2019) to run a linear mixed effects analysis. This allowed us to examine both within-subjects (celebrity status and deepfake pornography use) and between-subjects (sexuality: androphilic vs. gynephilic) factors' effects on judgements, as well as controlling for BJW and psychopathic traits. In this analysis, we used the 'participants' variable as a random intercept. The GAMLj module fit the full model as it was entered into the analysis within jamovi. All variables are mean-centred as a default within this software in order to make interpretation of the results easier to conduct. All *p*-values are corrected using the Bonferroni method. Descriptive statistics and internal consistency coefficients for all outcome measures are reported in Table 7.

In the model we entered each independent variable (vignette celebrity status: celebrity, friend, stranger; deepfake pornography use: own gratification vs. sharing; sexuality group: androphilic vs. gynephilic) as fixed factors, with participants' sex, age, relationship status, and their scores on the BJW and SRP4 scales as covariates. We fit this model separately for each outcome variable: perceptions of victim blame, criminality, and victim harm, and participants' self-reported proclivity to engage in deepfake pornography production.

Table 7. Descriptive statistics for all judgements, by condition and sexuality group

Victim	Use	Participant Sexuality	Victim blame			Perceived criminality			Victim harm			Proclivity		
Celebrity	Personal	Androphilic	1.64	± 0.71	(.38)	4.18	± 2.11	(.95)	4.37	± 1.99	(.97)	8.02	± 20.61	(-)
		Gynephilic	1.74	± 0.80	(.53)	3.70	± 2.06	(.96)	4.30	± 1.84	(.96)	12.75	± 23.73	(-)
	Sharing	Androphilic	1.58	± 0.81	(.58)	5.16	± 1.76	(.91)	4.85	± 1.79	(.96)	5.06	± 15.94	(-)
		Gynephilic	1.66	± 0.78	(.60)	4.64	± 1.89	(.94)	4.89	± 1.60	(.95)	8.25	± 18.73	(-)
Friend	Personal	Androphilic	1.60	± 0.91	(.70)	4.44	± 2.00	(.95)	5.49	± 1.67	(.95)	6.82	± 18.24	(-)
		Gynephilic	1.78	± 0.98	(.70)	3.81	± 2.03	(.94)	5.33	± 1.73	(.96)	10.64	± 21.59	(-)
	Sharing	Androphilic	1.44	± 0.78	(.66)	5.47	± 1.71	(.93)	6.03	± 1.26	(.94)	4.83	± 15.41	(-)
		Gynephilic	1.68	± 0.94	(.70)	4.98	± 1.81	(.92)	5.94	± 1.18	(.94)	5.68	± 15.32	(-)
Stranger	Personal	Androphilic	1.67	± 0.84	(.56)	4.52	± 1.96	(.95)	5.18	± 1.82	(.96)	7.69	± 19.57	(-)
		Gynephilic	1.84	± 0.95	(.64)	3.98	± 2.00	(.95)	5.05	± 1.79	(.96)	11.02	± 21.63	(-)
	Sharing	Androphilic	1.60	± 0.94	(.67)	5.68	± 1.43	(.89)	6.16	± 1.01	(.90)	5.71	± 18.14	(-)
		Gynephilic	1.78	± 0.98	(.68)	5.05	± 1.79	(.93)	5.90	± 1.15	(.90)	6.48	± 18.37	(-)

Note. Values represent mean averages ± 1 *SD* (Cronbach's α in parentheses)

Victim Blame Judgements

The linear mixed model explaining victim blame judgements fit the data well, with $R^2_{GLMM(m)} = .130$ and $R^2_{GLMM(c)} = .586$. Within the model there were significant effects of participants' age ($F(1, 351) = 4.60$, $p = .033$), belief in a just world ($F(1, 351) = 7.84$, $p = .005$), and psychopathic traits in relation to both lifestyle ($F(1, 351) = 5.52$, $p = .019$) and antisocial ($F(1, 351) = 40.86$, $p < .001$) domains of the SRP4. In relation to our experimental factors, there were main effects of the victims' celebrity status ($F(2, 1792) = 5.98$, $p = .003$) and the use of deepfake pornography ($F(1, 1792) = 12.19$, $p < .001$). There were no significant interactions in the model (Table 8).

Examining the estimates within the model, older age, a higher belief in a just world, lower lifestyle psychopathy, and higher antisocial psychopathy all predicted higher levels of victim blaming. The effect of celebrity status was accounted for by higher levels of victim blame being attributed in cases involving stranger victims (vs. celebrities, who acted as the reference category in the model). For use, the effect was driven by lower levels of victim blame being attributed where deepfake pornography was shared.

Table 8. Linear mixed effects model coefficients for victim blaming

			95% CI				
Variable	*Estimate*	*SE*	*Lower*	*Upper*	*df*	*t*	*p*
(Intercept)	**1.64**	**0.24**	**1.18**	**2.11**	**351**	**6.99**	**< .001**
Sex	-0.07	0.13	-0.33	0.19	352	-0.52	.603
Age	**0.01**	**0.01**	**0.00**	**0.02**	**351**	**2.15**	**.033**
Relationship status	0.09	0.08	-0.05	0.24	351	1.24	.215
Belief in a Just World	**0.02**	**0.01**	**0.01**	**0.03**	**351**	**2.80**	**.005**
SRP – Interpersonal	0.02	0.01	-0.00	0.04	351	1.89	.059
SRP – Affective	0.00	0.01	-0.02	0.03	351	0.16	.871
SRP – Lifestyle	**-0.02**	**0.01**	**-0.04**	**-0.00**	**351**	**-2.35**	**.019**
SRP – Antisocial	**0.09**	**0.01**	**0.06**	**0.12**	**351**	**6.39**	**< .001**
Friend victim	-0.03	0.03	-0.09	0.03	1792	-1.10	.272
Stranger victim	**0.07**	**0.03**	**0.01**	**0.13**	**1792**	**2.29**	**.022**
Use of deepfake pornography	**-0.09**	**0.03**	**-0.13**	**-0.04**	**1792**	**-3.49**	**< .001**
Sexuality group	0.03	0.13	-0.23	0.28	352	0.20	.844
Friend victim × Use of deepfake pornography	-0.07	0.06	-0.19	0.05	1792	-1.13	.257
Stranger victim × Use of deepfake pornography	0.01	0.06	-0.11	0.13	1792	0.17	.863
Friend victim × Sexuality group	**0.12**	**0.06**	**0.00**	**0.24**	**1792**	**2.02**	**.044**
Stranger victim × Sexuality group	0.09	0.06	-0.03	0.21	1792	1.53	.127
Use of deepfake pornography × Sexuality group	0.01	0.05	-0.08	0.11	1792	0.29	.776
Friend victim × Use of deepfake pornography × Sexuality group	0.09	0.12	-0.15	0.32	1792	0.72	.472
Stranger victim × Use of deepfake pornography × Sexuality group	0.05	0.12	-0.19	0.28	1792	0.38	.703
ICC				.525			
Observations				2163			
Marginal R^2 / Conditional R^2				.130 / .586			

Note. 'Victim' group estimates use the value 'Celebrity' as a reference category

Criminality Perceptions

The model explaining criminality judgements fit the data well, with $R^2_{GLMM(m)} = .198$ and $R^2_{GLMM(c)} = .753$. Within this model there were significant effects of participants' sex ($F(1, 351) = 15.36$, $p < .001$), age ($F(1, 351) = 7.52$, $p = .006$), belief in a just world ($F(1, 351)$ and levels of interpersonal psychopathy ($F(1, 351) = 13.33$, $p < .001$). In relation to the specific vignettes, there were main effects related to victim celebrity status ($F(2, 1794) = 29.38$, $p < .001$) and the use of deepfake pornography ($F(1, 1794) = 608.80$, $p < .001$). There was also a significant effect of sexuality group ($F(1, 351) = 4.26$, $p = .040$), but there were no significant interactions within the model (Table 9).

An analysis of these significant effects using the regression estimates reveals that women and younger participants were more likely to see the production of deepfake pornography as a criminal offence. Higher levels of both belief in a just world and interpersonal psychopathy were associated with reduced criminality judgements, meaning that these traits may be predictive of greater leniency in deepfake pornography cases. There were significant differences between all 'celebrity status' groups, with producing deepfake pornography of a stranger being viewed as the most criminal act, followed by producing material depicting a friend, and finally celebrity-based deepfake pornography being seen as the least criminal version of this behaviour. In relation to the use of deepfake pornography, criminality judgements were higher when the media were subsequently shared than when it was only used for personal sexual gratification. Of interest, gynephilic individuals (heterosexual men and homosexual women) were more likely to see deepfake pornography production as a criminal offence than were androphilic individuals (heterosexual women and homosexual men).

Table 9. Linear mixed effects model coefficients for perceived criminality

			95% CI				
Variable	*Estimate*	*SE*	*Lower*	*Upper*	*df*	*t*	*p*
(Intercept)	**2.70**	**0.56**	**1.60**	**3.81**	**351**	**4.79**	**< .001**
Sex	**1.25**	**0.32**	**0.62**	**1.87**	**351**	**3.92**	**< .001**
Age	**-0.04**	**0.01**	**-0.06**	**-0.01**	**351**	**-2.74**	**.006**
Relationship status	0.04	0.18	-0.31	0.39	351	0.21	.832
Belief in a Just World	**-0.03**	**0.01**	**-0.06**	**-0.01**	**351**	**-2.53**	**.012**
SRP – Interpersonal	**-0.09**	**0.02**	**-0.14**	**-0.04**	**351**	**-3.65**	**< .001**
SRP – Affective	0.00	0.03	-0.06	0.05	351	-0.13	.895
SRP – Lifestyle	-0.02	0.02	-0.07	0.02	351	-0.95	.343
SRP – Antisocial	0.06	0.03	-0.01	0.12	351	1.70	.089
Friend victim	**0.27**	**0.05**	**0.16**	**0.37**	**1794**	**5.08**	**< .001**
Stranger victim	**0.39**	**0.05**	**0.29**	**0.50**	**1794**	**7.51**	**< .001**
Use of deepfake pornography	**1.05**	**0.04**	**0.97**	**1.14**	**1794**	**24.67**	**< .001**
Sexuality group	**0.65**	**0.32**	**0.03**	**1.27**	**351**	**2.07**	**.040**
Friend victim × Use of deepfake pornography	0.13	0.10	-0.07	0.34	1794	1.28	.202
Stranger victim × Use of deepfake pornography	0.15	0.10	-0.05	0.36	1794	1.44	.149
Friend victim × Sexuality group	-0.05	0.10	-0.26	0.15	1794	-0.48	.630
Stranger victim × Sexuality group	-0.09	0.10	-0.29	0.12	1794	-0.82	.413
Use of deepfake pornography × Sexuality group	0.01	0.09	-0.16	0.18	1794	0.13	.896
Friend victim × Use of deepfake pornography × Sexuality group	0.17	0.21	-0.24	0.58	1794	0.82	.414
Stranger victim × Use of deepfake pornography × Sexuality group	-0.05	0.21	-0.46	0.36	1794	-0.24	.813
ICC				.692			
Observations				2165			
Marginal R^2 / Conditional R^2				.198 / .753			

Note. 'Victim' group estimates use the value 'Celebrity' as a reference category

Victim Harm Perceptions

The model explaining perceived victim harm explained a substantial proportion of the variance in this outcome, with $R^2_{GLMM(m)} = .185$ and $R^2_{GLMM(c)} = .636$. Within the model there were significant effects of participants' sex ($F(1, 351) = 3.91$, $p = .049$), age ($F(1, 351) = 14.95$, $p < .001$), and levels of psychopathy in the interpersonal domain ($F(1, 351) = 7.94$, $p = .005$). In relation to the manipulations that we made, there were main effects of the victims' celebrity status ($F(2, 1793) = 243.28$, $p < .001$) and the use of deepfake pornography ($F(1, 1793) = 224.93$, $p < .001$). There was also a significant interaction between these two factors ($F(2, 1793) = 6.99$, $p < .001$). Model coefficients are presented in Table 10.

Examining the coefficients within the model, we see that women and younger participants perceived higher levels of victim harm than men and older participants, respectively. As would be expected, lower levels of interpersonal psychopathy were also associated with higher levels of perceived victim harm. When comparing 'celebrity status' groups, less victim harm was perceived in cases involving celebrities than those involving friends or strangers. In relation to the eventual use of deepfake pornography, participants perceived higher levels of victim harm when the media is shared than when it is used for personal sexual gratification. The interaction between these variables (Figure 5) shows that the independent effects of use (personal sexual gratification vs. sharing) was consistent in cases involving victims who were either celebrities ($M_{diff} = 0.54$, $p < .001$, $d_z = 0.37$) or friends ($M_{diff} = 0.57$, $p < .001$, $d_z = 0.87$). However, when the victim was a stranger there was a larger effect of 'use', whereby less harm was perceived when the deepfake pornography was only used for personal sexual gratification ($M_{diff} = 0.90$, $p < .001$, $d_z = 0.61$).

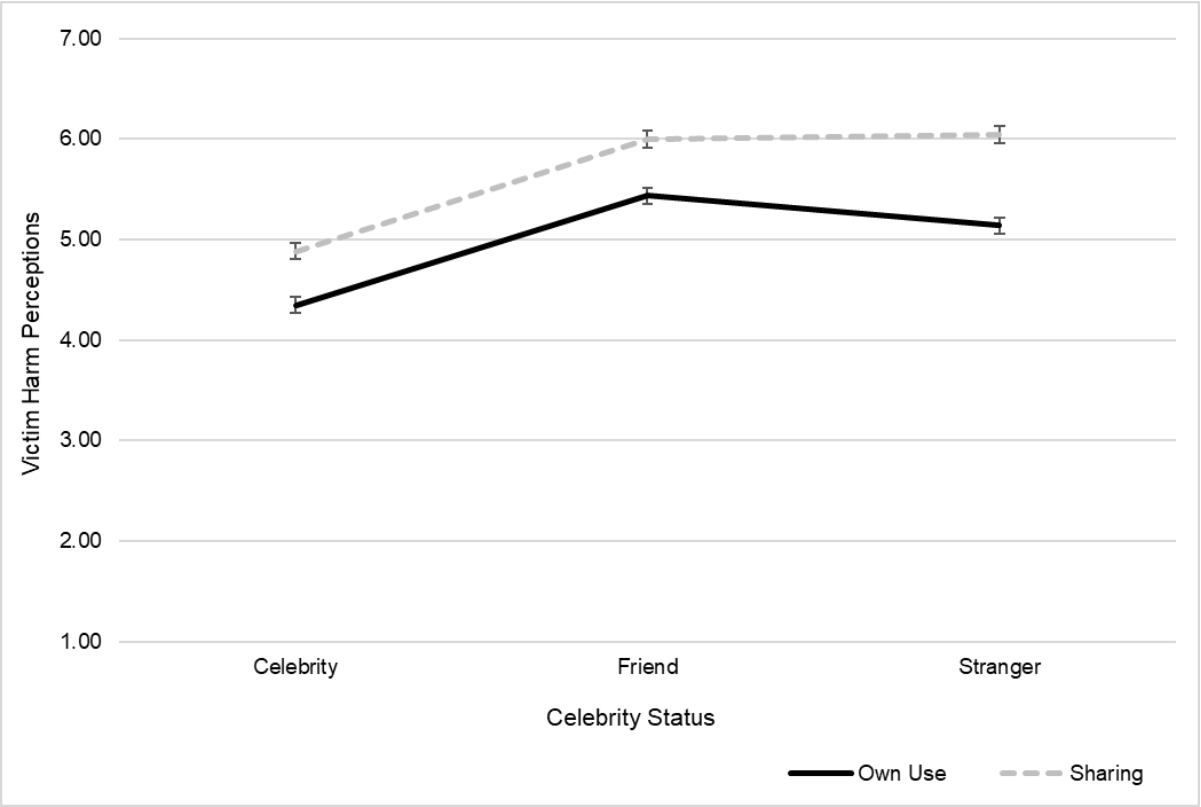

Figure 5. Celebrity Status × Deepfake Pornography Use interaction, predicting perceived victim harm

Table 10. Linear mixed effects model coefficients for perceptions of victim harm

			95% CI				
Variable	*Estimate*	*SE*	*Lower*	*Upper*	*df*	*t*	*p*
(Intercept)	**5.29**	**0.06**	**5.16**	**5.42**	**351**	**81.86**	**< .001**
Sex	**0.50**	**0.25**	**0.00**	**1.00**	**351**	**1.98**	**0.049**
Age	**-0.04**	**0.01**	**-0.06**	**-0.02**	**351**	**-3.87**	**< .001**
Relationship status	0.14	0.14	-0.14	0.42	351	1.01	0.314
Belief in a Just World	-0.02	0.01	-0.04	0.00	351	-1.93	0.055
SRP – Interpersonal	**-0.06**	**0.02**	**-0.09**	**-0.02**	**351**	**-2.82**	**0.005**
SRP – Affective	0.00	0.02	-0.04	0.05	351	0.17	0.868
SRP – Lifestyle	-0.01	0.02	-0.04	0.03	351	-0.33	0.74
SRP – Antisocial	0.00	0.03	-0.05	0.05	351	-0.11	0.914
Friend victim	**1.10**	**0.05**	**0.99**	**1.21**	**1793**	**20.16**	**< .001**
Stranger victim	**0.97**	**0.05**	**0.87**	**1.08**	**1793**	**17.83**	**< .001**
Use of deepfake pornography	**0.67**	**0.04**	**0.58**	**0.76**	**1793**	**15.00**	**< .001**
Sexuality group	0.38	0.25	-0.11	0.87	351	1.52	0.131
Friend victim × Use of deepfake pornography	0.03	0.11	-0.19	0.24	1793	0.26	0.795
Stranger victim × Use of deepfake pornography	**0.37**	**0.11**	**0.15**	**0.58**	**1793**	**3.36**	**< .001**
Friend victim × Sexuality group	-0.10	0.11	-0.31	0.12	1793	-0.89	0.372
Stranger victim × Sexuality group	-0.19	0.11	-0.40	0.03	1793	-1.73	0.084
Use of deepfake pornography × Sexuality group	0.03	0.09	-0.14	0.21	1793	0.35	0.724
Friend victim × Use of deepfake pornography × Sexuality group	-0.04	0.22	-0.46	0.39	1793	-0.17	0.866
Stranger victim × Use of deepfake pornography × Sexuality group	**-0.23**	**0.22**	**-0.66**	**0.20**	**1793**	**-1.05**	**0.295**
ICC				.553			
Observations				2164			
Marginal R^2 / Conditional R^2				.185 / .636			

Note. 'Victim' group estimates use the value 'Celebrity' as a reference category

Self-Reported Proclivity

The model for proclivity fit the data well, with $R^2_{GLMM(m)} = .099$ and $R^2_{GLMM(c)} = .537$. Compared to the previous outcomes there were fewer significant coefficients in relation to specific variables (Table 11). However, there were significant effects related to antisocial psychopathy ($F(1, 351) = 15.80$, $p < .001$) and the proposed use of deepfake pornography ($F(1, 1787) = 35.15$, $p < .001$). There was also a significant interaction between this 'use' variable and participants' sexuality grouping ($F(1, 1793) = 5.30$, $p = .021$). Model coefficients are presented in Table 11.

Examining the model coefficients, we see that higher levels of antisocial psychopathy is associated with a greater self-reported proclivity to engage in deepfake pornography production. In relation to the function of deepfake pornography, participants were more likely to express a proclivity to create such media for their own personal sexual gratification than to share it with others. Examining the interaction between this variable and sexuality (Figure 6), we found that androphilic participants (those attracted to men) did not differ in their self-reported proclivity to produce deepfake pornography for personal sexual gratification or for sharing ($M_{diff} = 2.05$, $p = .064$, $d_z = 0.13$). However, gynephilic participants (those attracted to women) reported being likely to produce deepfake pornography and subsequently share it than to use it for their own sexual gratification ($M_{diff} = 4.66$, $p < .001$, $d_z = 0.31$).

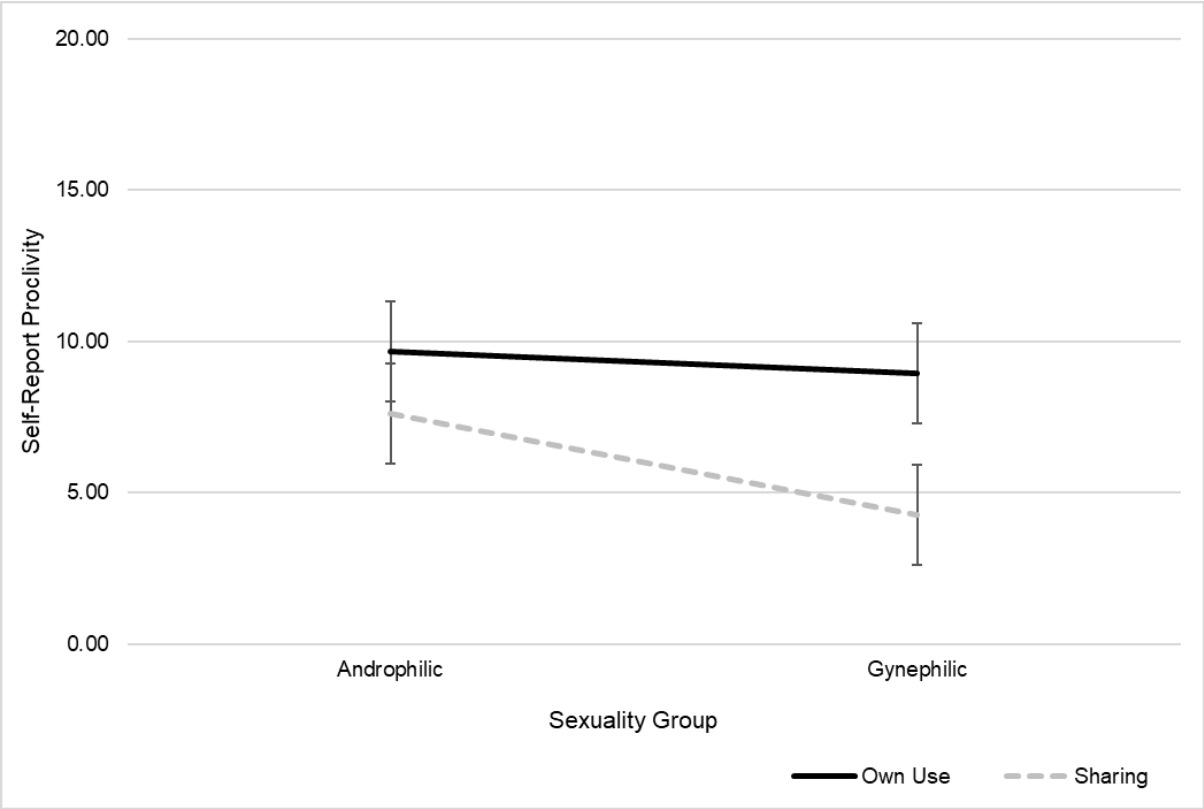

Figure 6. Deepfake Pornography Use × Sexuality Group interaction, predicting self-reported proclivity

Table 11. Linear mixed effects model coefficients for self-reported proclivity

			95% CI				
Variable	*Estimate*	*SE*	*Lower*	*Upper*	*df*	*t*	*p*
(Intercept)	**7.72**	**0.73**	**6.29**	**9.15**	**351**	**10.57**	**< .001**
Sex	-3.08	2.86	-8.68	2.52	351	-1.08	.282
Age	-0.05	0.12	-0.28	0.18	351	-0.42	.674
Relationship status	-0.82	1.61	-3.97	2.32	351	-0.51	.608
Belief in a Just World	0.21	0.12	-0.02	0.44	352	1.78	.076
SRP – Interpersonal	0.29	0.22	-0.15	0.72	351	1.30	.196
SRP – Affective	0.09	0.26	-0.42	0.59	352	0.33	.738
SRP – Lifestyle	0.16	0.22	-0.27	0.59	352	0.71	.475
SRP – Antisocial	**1.18**	**0.30**	**0.60**	**1.76**	**351**	**3.97**	**< .001**
Friend victim	**-1.44**	**0.69**	**-2.80**	**-0.08**	**1786**	**-2.07**	**.038**
Stranger victim	-0.80	0.69	-2.16	0.57	1787	-1.15	.252
Use of deepfake pornography	**-3.36**	**0.57**	**-4.47**	**-2.25**	**1787**	**-5.93**	**< .001**
Sexuality group	-2.02	2.84	-7.59	3.54	351	-0.71	.476
Friend victim × Use of deepfake pornography	0.10	1.39	-2.62	2.82	1786	0.07	.941
Stranger victim × Use of deepfake pornography	0.50	1.39	-2.22	3.22	1787	0.36	.720
Friend victim × Sexuality group	-1.88	1.39	-4.60	0.84	1786	-1.35	.176
Stranger victim × Sexuality group	-1.91	1.39	-4.63	0.81	1787	-1.38	.168
Use of deepfake pornography × Sexuality group	**-2.61**	**1.13**	**-4.83**	**-0.39**	**1787**	**-2.30**	**.021**
Friend victim × Use of deepfake pornography × Sexuality group	-1.09	2.77	-6.53	4.35	1786	-0.39	.695
Stranger victim × Use of deepfake pornography × Sexuality group	-1.17	2.78	-6.61	4.27	1787	-0.42	.673
ICC				.486			
Observations				2156			
Marginal R^2 / Conditional R^2				.099 / .537			

Note. 'Victim' group estimates use the value 'Celebrity' as a reference category

Knowledge of the Deepfake Pornography Label

As in Study 1, we coded participants' labelling of the behaviour depicted in the scenarios as 'correct' if it included the word 'deepfake'. Of the 364 participants in the sample, 16 correctly identified the behaviour. This represents a 4.4% accuracy rate, which is comparable to that reported in Study 1 (6.6%).

General Discussion

In two independently sampled studies, we consistently report low public awareness of deepfake media production, and more lenient judgements of deepfaking when victims were celebrities (relative to non-celebrities) and/or males (relative to females). Moreover, more lenient judgements were also predicted by variation (i.e., increased) in self-reported psychopathy and belief in a just world.

Consistent across both samples, participants deemed incidents of deepfakes involving celebrity victims to be less criminal and less harmful on average, relative to non-celebrity victims. Specifically, evidence from Study 2 suggests this disparity exists when comparing against both victims who are complete strangers, and victims who are friends of the perpetrator. This evidence is of particular concern given that the gross proportion of deepfakes that are generated and disseminated feature female celebrities (Citron & Chesney, 2019; Delfino, 2019); potentially as a functiopn of those featuring non-celebrities conveying little market value. Finance was reported as the primary motivating factor in recent arrests for deepfaking (Japan Gazette, 2020). It is clear from the *online trolling* literature that celebrities are the predominant targets of online abuse (Garde-Hansen & Gorton, 2013), with said victims internalising the detrimental impact of such behaviour (Ouvrein et al., under submission). Seemingly, our data mirrors that of other investigations into the perceived impact of internet-mediated abuse (Ouvrein et al., 2017; Ouvrein et al., 2018).

Absent from this data, however, is the rationale as to *why* our participants responded in this way. Previous literature has indicated that lower perceptions of criminality and harm may be attributed to considerations that abuse is part and parcel of being famous (Ouvrein et al., 2017; Ouvrein et al., 2018), that such behaviour is safer against celebrities (Feasey, 2008), or that celebrities have somehow brought said behaviour upon themselves (Scott et al., 2018; 2020). As future research seeks to delineate this further, it should bear in mind the potential role of participant sex differences in said rationale. Specifically, our data indicated that on average, male participants reported greater blame for celebrity victims, with female participants reporting greater blame for non-celebrity victims.

The role of sex was also applicable in the context of the victim. In Study 1, vignettes featuring female victims predicted less victim blame and greater levels of harm and criminality when compared to vignettes featuring male victims. In practice, this suggests that the experiences of male victims of deepfakes may be viewed less seriously than their female counterparts, however there is no empirical research to support this on an experiential level. Of interest and to the best of our knowledge, there are very few incidents of high-profile male victims of deepfaking and no documented accounts of their experiences and associated impact. However, this is not to say that this group of individuals do not exist. We know males become victims of other image-based sexual offences such as revenge pornography (Hall & Hearn, 2018), wherein they can suffer psychological impacts including depression, anxiety, and stress related to trust and self-image (Bates, 2017). Drawing on the wider sexual abuse literature, the low visibility of male victims might be somewhat accounted for by traditional gender norms (e.g., being invulnerable) and fear of disbelief concealing the disclosure of victimisation (Spiegel, 2003; von Hohendorff et al., 2017). Relating to our research, it is possible that the patterns of judgements disseminated here might further add to a lack of disclosure (e.g., feeling their experiences are not as important compared to female victims), and as such we fully endorse future applied work which both [1] identifies and tries to better understand male victims of deepfaking, and [2] reduces their barriers to disclosing and seeking help.

Regarding proclivity metrics, in Study 1 our participants expressed a greater proclivity on average towards *creating* deepfakes involving celebrities relative to non-celebrities, however this disparity did not extend to the *sharing* of deepfakes. Overall, men expressed higher proclivity scores than females. To help understand this, Study 2 extended this data through the assessment of image 'use'. Participants perceived greater victim harm when the images were shared (relative to when they were used for personal sexual gratification) for both celebrity and friend victims. However, when the victim was a stranger, there was a further reduction in perceived harm when the media was used for personal sexual gratification. This lends some support to the theoretical work of both Harper et al. (2021) and Harris (2019), where they hypothesise that the engagement in deepfake media production might include curiosity, sexual compulsivity, or a specific sexual interest. As such, engagement in the creation of dissemination of deepfakes might reflect the dysfunctional formation of sexual scripts (Wright, 2014) denoting healthy sexual behaviour, combined with sexual desires to see intimate images of desired others (Harper et al., 2021).

Moreover, both the interpersonal and antisocial facets of psychopathy predicted proclivity to generate and share deepfakes across both of our studies, with the antisocial facet predicting greater victim blame and reduced criminality and victim harm as a result. This association was predicted, and maps directly onto both our existing knowledge that psychopathy broadly predicts antisocial behaviour (Blais et al., 2014; Marsh & Cardinale, 2012), and especially relevant to this research area, online aggression, trolling, and proclivity to sexually harass (Buckels et al., 2014; Pabian et al., 2015; Sumner et al., 2012; Zeigler-Hill et al., 2016) as well as image-based sexual offending (i.e., revenge pornography) more specifically (Pina et al., 2017). The precise mechanism underpinning this association is likely to be a function of those scoring higher on psychopathy having reduced empathy for the impact of others (Viding & McCory, 2019) combined with gaining pleasure from inflicting emotional distress on others through their actions (Harper et al., 2021; Kircaburun et al., 2018; Sest & March, 2017). Belief in a just world also, and naturally, predicted victim blame; supporting qualitative claims about the role that the victim plays in their own victimisation in sexual abuse (Vonderhaar & Carmody, 2015) and image-based sexual offence studies more broadly (Henry et al., 2017; Henry & Powell, 2015). However, it remains unknown as to *how* and in what specific ways our participants rationalised the contribution of the victim to becoming the subject of deepfaking, or whether this position was even a conscious one.

Another interesting finding that was unexpected in our data was that although the two sexuality groups in Study 2 did not differ in their proclivity for producing deepfake pornography, gynephilic individuals (typically heterosexual men, but a group that also includes homosexual women) were less likely to share such material. This is in direct opposition to what might be expected when considering image-based sexual abuse (of which deepfaking is a constituent behavioural pattern) as a gendered crime motivated by a desire to exert patriarchal power and control (see McGlynn et al., 2017). Instead, this finding is consistent with the view that (at least some) image-based sexual abuse offences are driven by a desire for personal sexual gratification (Fido & Harper, 2020). That is, while women (typically) are equally likely to both produce and share deepfake pornography (indicating more social motivations), for men this behaviour may be more driven by personal desires, leading to the combination of an equal proclivity to women in terms of deepfake production, but a lower proclivity towards sharing.

A final important finding of our research was how few participants, across either sample (6.6%, 4.4%), were able to accurately name (or get close to naming) the action of creating and disseminating deepfake pornography. Although this is a predictable finding,

given that compared to revenge pornography and upskirting, deepfaking remains a relatively new type of image-based sexual offence that has received little media attention (Cole, 2018), victims have included high-profile names (Delfino, 2019). Given the widely understood social and personal implications of becoming a victim of image-based sexual offending (Bates, 2017; Bloom, 2014), it is important that potential victims are able to define and articulate what has happened to them; especially if there is to be a movement towards further criminalising such actions Worldwide. Future work might attempt to generate knowledge and understanding via (social)media posts and articles, which might also carry the message of the damage that such behaviour has the potential to cause; thus helping to attenuate future instances.

Limitations and Future Directions

First, both studies within this investigation sampled from populations within the UK. The rational for this was to ensure contextual consistency for our responders without having to control for factors such as variation in legislation, which although is lacking Worldwide, has been shown to vary greatly across American states (Delfino, 2019; Harris, 2019). As such, although our findings should still be considered to have somewhat of an international impact due to overall lack of legislation pertaining to deepfake pornography, we do acknowledge some of the potential cultural differences (e.g., values and norms) which might impact the findings observed here (Fido & Harper, 2020). Second, the vignettes used within this investigation would benefit from further validation. Although we are confident in the accurate depiction and description of instances of deepfake pornography generation and dissemination, future research would benefit from liaising with individuals with lived experiences to ensure accuracy, both from an ethical perspective and from the standpoint of scientific validity. Third, it would be remiss of us to not recognise that the sequence of the presentation of within-subject tasks and materials might have had – an albeit small – impact on the observed results. We hope to have alleviated this to some extend through the method of randomising the order of presentation. Fourth, despite a firm theoretical grounding being offered for the inclusion of psychopathy as a co-variate of interest, this might be a relatively reductionist approach owing to literature in the field documenting the role of allied *dark* traits such as narcissism and Machiavellianism in predicting the endorsement of unprovoked celebrity- and non-celebrity-focused aggression and victim blame (Scott et al., 2020). Future research might invite measurement of these facets in order to broaden our understanding of their role in deepfake media production and dissemination. Finally, although proclivity data

was collected, we are yet to understand *why* such individuals would want to generate and/or disseminate such media. This is an important future step because such rationale might be able to feed into advisory information and intervention programmes to prevent future deepfaking instances.

Conclusion

To conclude, this paper is the first to examine judgements of deepfake media production and the predictors of one's proclivity to generate and disseminate such images. It builds on theoretical research (Harper et al., 2021), and in the context of the first Worldwide arrests for such behaviour (Japan Gazette, 2020) presents a timely investigation. Our data suggests the presence of generally low UK public awareness and more lenient judgements towards celebrity and male victims, which are proliferated through the presence of greater self-report psychopathy and beliefs about a just world. Our inferences suggest a need for a better understanding about victims of deepfake media production, and rationales pertaining to *why* an individual might choose to generate and disseminate such images. Together, we hope this data provides a useful first step to support the generation of means to protect against such behaviour, with future research tasked with identifying and reducing barriers to victim disclosure.

References

Ayala, E. E., Kotary, B., & Hetz, M. (2018). Blame Attributions of Victims and Perpetrators: Effects of Victim Gender, Perpetrator Gender, and Relationship. *Journal of Interpersonal Violence, 33*(1), 94–116. https://doi.org/10.1177/0886260515599160

Abusive Behaviour and Sexual Harm Act (Scotland). Retrieved 8 September 2019, from http://www.legislation.gov.uk

Blair, R. J. R., Monson, J., & Frederickson, N. (2001). Moral reasoning and conduct problems in children with emotional and behavioural difficulties. *Personality and Individual Differences, 31*(5), 799–811. https://doi.org/10.1016/S0191-8869(00)00181-1

Blair, R.J. (1995). A cognitive developmental approach to morality: investigating the psychopath. *Cognition, 57*(1), 1–29. https://doi.org/10.1016/0010-0277(95)00676-p

Bohner, G., Siebler, F., & Schmelcher, J. (2006). Social Norms and the Likelihood of Raping: Perceived Rape Myth Acceptance of Others Affects Men's Rape Proclivity. *Personality and Social Psychology Bulletin, 32*(3), 286–297. https://doi.org/10.1177/0146167205280912

Brewer, G., Hunt, D., James, G., & Abell, L. (2015). Dark Triad traits, infidelity and romantic revenge. *Personality And Individual Differences, 83*, 122-127. doi: 10.1016/j.paid.2015.04.007

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences, 67*, 97–102.

Burkell, J., & Gosse, C. (2019). Nothing new here: Emphasizing the social and cultural context of deepfakes. *First Monday, 24* (12). doi: http://dx.doi.org/10.5210/fm.v24i12.10287

Burt, M. R. (1980). Cultural myths and supports for rape. *Journal of Personality and Social Psychology, 38*(2), 217–230. https://doi.org/10.1037/0022-3514.38.2.217

Bustamante, A., & Chaux, E. (2014). Reducing moral disengagement mechanisms: A comparison of two interventions. Journal of Latino-Latin American Studies, 6, 52–54

Cardinale, E.M. & Marsh, A.A. (2015) Impact of Psychopathy on Moral Judgments about Causing Fear and Physical Harm. *PLoS ONE 10*(5). https://doi.org/10.1371/journal.pone.0125708

Chesney, R. & Citron, D. (2019). Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics Essays. *Foreign Affairs. https://alinstitute.org/images/Library/DeepfakesAndDisinformationWar.pdf*

Citron, D.K. & Chesney, R. (2019) Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security.*107 California Law Review 1753.* University of California Berkeley School of Law. https://scholarship.law.bu.edu/faculty_scholarship/640

Coakes, S.J., (2005). *SPSS: Analysis without Anguish-version 12.0 for Windows.* John Wiley & Sons Ltd.

Cohen J. E. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc

Cole, S. (2017, December 11). AI-assisted fake porn is here and we're all fucked. *Motherboard Tech by Vice.* https://motherboard.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn

Combs-Lane, A. M., & Smith, D. W. (2002). Risk of Sexual Victimization in College Women: The Role of Behavioral Intentions and Risk-Taking Behaviors. *Journal of Interpersonal Violence, 17*(2), 165–183. https://doi.org/10.1177/0886260502017002004

Cracker, N. & March, E. (2016). The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences, 102*, 79-84. https://www.sciencedirect.com/science/article/abs/pii/S0191886916307930?via%3Dihub

Dalbert, C. (2009). Belief in a just world. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of Individual Differences in Social Behavior,* 288-297. New York: Guilford Publications.

Davis, K. C., Norris, J., George, W. H., Martell, J., & Heiman, J. R. (2006). Rape-Myth Congruent Beliefs in Women Resulting from Exposure to Violent Pornography: Effects of Alcohol and Sexual Arousal. *Journal of Interpersonal Violence, 21*(9), 1208–1223. https://doi.org/10.1177/0886260506290428

Dodge, A. (2016). Digitizing rape culture: Online sexual violence and the power of the digital photograph. *Crime, Media, Culture, 12*(1), 65-82. https://doi.org/10.1177/1741659015601173

Dooley, J., Pyzalski, J., & Cross, D. (2009). Cyberbullying versus face-to-face bullying: A theoretical and conceptual review. *Journal of Psychology, 217,* 182–188.

Durall, R., Keuper, M., Pfreundt, F.J., & Keuper, J. (2019). Unmasking DeepFakes with simple Features. *Cornell University*. https://arxiv.org/abs/1911.00686

Eaton, A.A., Jacobs, H. & Ruvalcaba, Y. (2017). 2017 Nationwide Online Study of Nonconsensual Porn Victimization and Perpetration - A Summary Report. *Cyber Civil Rights Initiative.* Florida International University, Department of Psychology

https://www.cybercivilrights.org/wp-content/uploads/2017/06/CCRI-2017-Research-Report.pdf

Farokhmanesh, M. (2018, January 30). Is it legal to swap someone's face into porn without consent? *The Verge.* https://www.theverge.com/2018/1/30/16945494/deepfakes-porn-face-swap-legal

Feasey, R. (2008). Reading Heat: The meanings and pleasures of star fashions and celebrity gossip. *Continuum: Journal of Media & Cultural Studies, 22*(5), 687–699.

Fido, D., Harper, C. A., Davis, M. A., Petronzi, D., & Worrall, S. (2021). Intrasexual Competition as a Predictor of Women's Judgments of Revenge Pornography Offending. *Sexual Abuse, 33*(3), 295-320. https://doi.org/10.1177/1079063219894306

Field, A. P. (2007). *Discovering statistics using SPSS*. [electronic resource]. SAGE Publications.

Floridi, L. (2018) Artificial Intelligence, Deepfakes and a Future of Ectypes. *Philosophy & Technology, 31*(3) 317-321. https://doi.org/10.1007/s13347-018-0325-3

Franks, M.A. & Waldman, A.E. (2020). Sex, Lies and Videotape: Deep Frakes and Free Speech Delusions. *Maryland Law Review*. *78*(4), 893. https://digitalcommons.law.umaryland.edu/cgi/viewcontent.cgi?article=3835&context=mlr

Gallucci, M. (2019). *GAMLj: General analyses for linear models*. [jamovi module].

Garde-Hansen, J., & Gorton, K. (2013). *Emotion online: Theorizing affect on the internet.* New York, NY: Palgrave Macmillan.

Gavin, J., & Scott, A. J. (2019). Attributions of victim responsibility in revenge pornography. *Journal of Aggression, Conflict and Peace Research, 1*1(4), 263-272. https://research.gold.ac.uk/26593/1/2019%20Gavin%20%26%20Scott%20%28JACPR%29.pdf

Ghanem, K. G., Hutton, H. E., Zenilman, J. M., Zimba, R., & Erbelding, E. J. (2005). Audio computer assisted self interview and face to face interview modes in assessing response bias among STD clinic patients. *Sexually Transmitted Infections, 81*, 421-425. https://sti.bmj.com/content/sextrans/81/5/421.full.pdf

Glenn, A. L., Iyer, R., Graham, J., Koleva, S.P., & Haidt, J. (2009). Are All Types of Morality Compromised in Psychopathy?. *Journal of personality disorders 23*(4):384-98. DOI: 10.1521/pedi.2009.23.4.384

Goodboy, A. K., & Martin, M. M. (2015). The personality profile of a cyberbully: Examining the Dark Triad. *Computers In Human Behavior, 49*, 1-4. doi: 10.1016/j.chb.2015.02.052

Green, S. B. (1991). How many subjects does it take to do a regression analysis? Multivariate Behavioral Research, 26, 499-510.

Grubb, A. R. & Turner, E. (2012). Attribution of blame in rape cases: a review of the impact of rape myth acceptance, gender role conformity and substance use on victim blaming. *Aggression and Violent Behaviour, 17*(5), 443-52. https://tinyurl.com/y8rgwmtq

Grubb, A. R., & Harrower, J. (2009). Understanding attribution of blame in cases of rape: An analysis of participant gender, type of rape and perceived similarity to the victim. *Journal Of Sexual Aggression, 5*(1), 63–81. https://www.tandfonline.com/doi/abs/10.1080/13552600802641649

Gurnham, D. (2016). A Critique of Carceral Feminist Arguments on Rape Myths and Sexual Scripts. *New Criminal Law Review*, *19*(2), 141-170. doi:10.1525/nclr.2016.19.2.141

Hale, W. (2019) First Federal Legislation on Deepfakes Signed Into Law. *JD Supra*. https://www.jdsupra.com/legalnews/first-federal-legislation-on-deepfakes-42346/

Hand, C J., Scott, G. G., Brodie, Z. P., Xilei, Y., & Sereno, S. C. (2021). Tweet valence, volume of abuse, and observers' Dark Tetrad personality factors influence victim-blaming and the perceived severity of Twitter cyberabuse. *Computers in Human Behavior Reports*, 100056. https://doi.org/10.1016/j.chbr.2021.100056

Harper, C. A., Smith, L., Leach, J., Daruwala, N., & Fido, D. (2020, November 28). Development and validation of the Beliefs about Revenge Pornography Questionnaire. https://doi.org/10.31234/osf.io/6qr7t

Harper, C. A., Fido, D., & Petronzi, D. (2021). Delineating non-consensual sexual image offending: Towards an empirical approach. *Aggression and Violent Behavior, 58*, 101547. https://doi.org/10.1016/j.avb.2021.101547

Harper, C. A., Smith, L., & Fido, D. *Development and validation of the Revenge porn Myth Scale* (in prep).

Harris, D. (2019). Deepfakes: False pornography is here and the law cannot protect you. *Duke Law & Technology Review,17*, 99-128. https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=1333&context=dltr

Hart, S. D. (1998). The role of psychopathy in assessing risk for violence: Conceptual and methodological issues. *Legal and Criminological Psychology*, *3*(1), 121-137. doi: 10.1111/j.2044-8333.1998.tb00354.x

Harwell, D. (2018, December 30). Fake-porn videos are being weaponized to harass and humiliate women: 'Everybody is a potential target'. *Washington Post.*

tps://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target/

Hayes, R. M., Lorenz, K., & Bell, K. A. (2013). Victim Blaming Others: Rape Myth Acceptance and the Just World Belief. *Feminist Criminology, 8*(3), 202–220. https://doi.org/10.1177/1557085113484788

Howitt, D., & Cramer, D. (2011). *Introduction to SPSS statistics in psychology : for version 19 and earlier* (5th ed.). Pearson Education M.U.A.

Hearn, J., & Hall, M. (2019). 'This is my cheating ex': Gender and sexuality in revenge porn. *Sexualities, 22*(5–6), 860–882. https://doi.org/10.1177/1363460718779965

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige. *Evolution and Human Behavior, 22,* 165–196.

Henry, N. & Powell, A. (2018). Technology-Facilitated Sexual Violence: A Literature Review of Empirical Research. *Trauma, Violence, & Abuse, 19*(2), 195-208.

Japan Gazette (October 2, 2020). Adult video creation with "Deep Fake" technology or first arrest. Japan Gazette. Retrieved from: https://japangazette.com/2020/10/02/adult-video-creation-with-deep-fake-technology-or-first-arrest/

Kietzmann, J., Lee, L., McCarthy, I., & Kietzmann, T. (2020). Deepfakes: Trick or treat? *Business Horizons, 63*(2), 135-146. doi: 10.1016/j.bushor.2019.11.006

Kircaburun, K., Demetrovics, Z., & Tosuntaş, Ş. B. (2018). Analyzing the links between problematic social media use, Dark Triad traits, and self-esteem. *International Journal of Mental Health and Addiction*. https://doi.org/10.1007/s11469-018-9900-1.

Kircaburun, K., Jonason, P. K. & Griffiths, M. D. (2018). The Dark Tetrad traits and problematic social media use: The mediating role of cyberbullying and cyberstalking. *Personality and Individual Differences, 135,* 264–269. https://doi.org/10.1016/j.paid.2018.07.034

Kokkinos, C., Antoniadou, N., & Markos, A. (2014). Cyber-bullying: An investigation of the psychological profile of university student participants. *Journal Of Applied Developmental Psychology, 35*(3), 204-214. doi: 10.1016/j.appdev.2014.04.001

Krahé, B., Temkin, J., & Bieneck, S. (2007). Schema-driven information processing in judgements about rape. *Applied Cognitive Psychology, 21*(5), 601-619. doi: 10.1002/acp.1297

Lerner, M., & Simmons, C. H. (1966). Observer's reaction to the 'innocent victim': Compassion or rejection? *Journal of Personality and Social Psychology, 4*(2), 203–210.

Li, Y., Yang, X., Sun, P., Qi, H. & Lyu, S. (2019). Celeb-DF: A New Dataset for Deepfake Forensics. arXiv preprint arXiv:1909.12962. https://arxiv.org/abs/1909.12962

Lipkusa, I. M., Dalbert, C., & Siegler, I. C. (1996). The Importance of Distinguishing the Belief in a Just World for Self Versus for Others: Implications for Psychological Well-Being. *Personality and Social Psychology Bulletin, 22*(7), 666–677. https://doi.org/10.1177/0146167296227002

Marsh, A. A. & Cardinale, E. M. (2012). Psychopathy and Fear: Specific Impairments in Judging Behaviours that Frighten Others. *Emotion*, 12, 892–8. American Psychological Association. https://pdfs.semanticscholar.org/40b8/bcaecb35e9ccd52869853b854a9d320111eb.pdf

Marsh, A. A., & Cardinale, M. E. (2014). When psychopathy impairs moral judgments: neural responses during judgments about causing fear. *Social Cognitive and Affective Neuroscience, 9*(1), 3–11, https://doi.org/10.1093/scan/nss097

McGlynn, C., & Rackley, E. (2017). Image Based Sexual Abuse. *Oxford Journal of Legal Studies, 37*(3), 534 – 561. doi:10.1093/ojls/gqw033

McGlynn, C., Rackley, E. & Houghton, R. (2017). Beyond 'Revenge Porn': The Continuum of Image-Based Sexual Abuse. *Fem Leg Stud 25*, 25–46. https://doi.org/10.1007/s10691-017-9343-2

McGregor, J. (2019, July 19). FaceApp: How does it profit from your data? Is it dangerous? *Forbes*. https://www.forbes.com/sites/jaymcgregor/2019/07/19/faceapp-how-does-it-profit-from-your-data-is-it-dangerous/#1750c5326f83

McLeod, S. A. (2019). Extraneous variables. *Simply Psychology*. https://www.simplypsychology.org/extraneous-variable.html

McMahon, S., & Farmer, G. (2011). An Updated Measure for Assessing Subtle Rape Myths. *Social Work Research, 35(2),* 71-81. www.jstor.org/stable/42659785

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of cardiac anaesthesia*, *22*(1), 67–72. https://doi.org/10.4103/aca.ACA_157_18

Morgan, G. A. (2004). *SPSS for Introductory Statistics : Use and Interpretation* (2nd ed.). [electronic resource]. Taylor & Francis [CAM].

Muntean, N., & Petersen, A. H. (2009). Celebrity twitter: Strategies of intrusion and disclosure in the age of technoculture. M/C Journal: *A Journal of Media & Culture, 12*(5).

Ng, A. (2019, December 2). FBI calls FaceApp a 'potential counterintelligence threat' from Russia. *Cnet.* https://www.cnet.com/news/fbi-calls-faceapp-a-potential-counterintelligence-threat-from-russia/

Ouvrein, G., Vandebosch, H., & De Backer, C. J. S. (2017). Celebrity critiquing: Hot or not? Teenage girls' attitudes and responses to the practice of negative celebrity critiquing. Celebrity Studies, 8(3), 461–476.

Ouvrein, G., Vandebosch, H., & De Backer, C .J. S. (under submission). The celebritycyberbullying experience and coping guide. A framing analysis of online celebritycyberbullying citations in news articles. In H. Vandebosch & L. Green (Eds.),

Pabian, S., De Backer, C. J. S., & Vandebosch, H. (2015). Dark Triad personality traits and adolescent cyber-aggression. *Personality and Individual Differences, 75*, 41–46.

Pabian, S., Vandebosch, H., Poels, K., Van Cleemput, K., & Bastiaensens, S. (2016). Exposure to cyberbullying as a bystander: An investigation of desensitization effects among early adolescents. *Computers in Human Behavior, 62*, 480–487

Pallant, J. (2010). *SPSS survival manual.: a step by step guide to data analysis using SPSS* (4th ed.). [electronic resource]. McGraw-Hill.

Patella-Rey, P. J. (2018) Beyond privacy: bodily integrity as an alternative framework for understanding non-consensual pornography. *Information, Communication & Society, 21*(5), 786-791. DOI: 10.1080/1369118X.2018.1428653

Paulhus, D. L., Neumann, C. S., & Hare, R. D. (2009). *Manual for the Self-Report Psychopathy Scale* (4th ed.). Toronto: Multi-Health Systems.

Payne, L. D., Lonsway, A. K., & Fitzgerald, F. L. (1999). Rape Myth Acceptence: Exploration of Its Structure and Its Measurement Using the Illinois Rape Myth Acceptence Scale. *Journal of Research in Personality, 33*(1), 27-68. https://doi.org/10.1006/jrpe.1998.2238

Peng, X., Li, Y., Wang, P., Mo, L., & Chen, Q. (2015). The ugly truth: Negative gossip about celebrities and positive gossip about self entertain people in different ways. Social Neuroscience, 10(3), 37–41.

Pina, A., Holland, J., & James, M. (2017). The Malevolent Side of Revenge Porn Proclivity. *International Journal of Technoethics, 8*(1), 30-43. https://doi.org/10.4018/ijt.2017010103

Pornari, C. D., & Wood, J. (2010). Peer and cyber aggression in secondary school students: The role of moral disengagement, hostile attribution bias, and outcome expectancies. *Aggressive Behavior, 36*, 81–94.

Powell, A., Henry, N., Flynn, A., & Scott, A. (2018). Image-based sexual abuse: The extent, nature, and predictors of perpetration in a community sample of Australian residents. *Computers in Human Behavior, 92*, 393-402. DOI: 10.1016/j.chb.2018.11.009

Predictor of Women's Judgments of Revenge Pornography Offending. *Sexual Abuse*, doi: 10.1177/1079063219894306

Rasmussen, K. R., & Boon, S. D. (2014). Romantic revenge and the Dark Triad: A model of 843 impellance and inhibition. *Personality and Individual Differences, 56*, 51-56. doi: 844 10.1016/j.paid.2013.08.018

Renati, R., Berrone, C., & Zanetti, M. A. (2012). Morally disengaged and unempathic: Do cyberbullies fit these difinitions? An exploratory study. Cyberpsychology, Behavior, and Social Networking, 15(8), 391–398.

Ritchie, M. B., Blais, J., Forth, A. E., & Book, A. S. (2018). Identifying vulnerability to violence: the role of psychopathy and gender. *Journal of Criminal Psychology, 8*(2), 125-137. https://doi.org/10.1108/JCP-06-2017-0029

Robertson, A. (2019, July 1). Virginia's 'revenge porn' laws now officially cover deepfakes. *The Verge*. https://www.theverge.com/2019/7/1/20677800/virginia-revenge-porn-deepfakes-nonconsensual-photos-videos-ban-goes-into-effect

Russell, K. J., & Hand, C. J. (2017). Rape myth acceptance, victim blame attribution and Just World Beliefs: A rapid evidence assessment. *Aggression And Violent Behavior, 37*, 153-160. doi: 10.1016/j.avb.2017.10.008

Scott, A. J., & Gavin, J. (2018). Revenge pornography: the influence of perpetrator-victim sex, observer sex and observer sexting experience on perceptions of seriousness and responsibility. *Journal of Criminal Psychology, 8* (2), 162-72.

Scott, G. G., Brodie, Z. P., Wilson, M. J., Ivory, L., Hand, C. J., & Sereno, S. C. (2020). Celebrity abuse on Twitter: The impact of tweet valence, volume of abuse, and dark triad personality factors on victim blaming and perceptions of severity. *Computers in Human Behavior, 103*, 109-119.

Scott, G. G., Wienercz, S., & Hand, C. J. (2018). The frequency and source of online abuse impacts attribution of victim blame and perceptions of victim attractiveness. *Computers in Human Behavior, 92*, 119–127.

Sest, N., & March, E. (2017). Constructing the cyber-troll: Psychopathy, sadism, and empathy. *Personality and Individual Differences, 119*, 69-72. doi: 10.1016/j.paid.2017.06.038

Sleath, E., & Bull, R. (2010). Male Rape Victim and Perpetrator Blaming. *Journal of Interpersonal Violence, 25*(6), 969–988. https://doi.org/10.1177/0886260509340534

Spiegel, J. (2003). *Sexual abuse of males: The SAM model of theory and practice*. New York: Routledge

Statt, N. (2019, November 29). China makes it a criminal offense to publish deepfakes or fake news without disclosure Mirroring new California law on political ads. *The Verge*. https://www.theverge.com/2019/11/29/20988363/china-deepfakes-ban-internet-rules-fake-news-disclosure-virtual-reality

Stubbs-Richardson, M., Rader, N. E., & Cosby, A. G. (2018). Tweeting rape culture: Examining portrayals of victim blaming in discussions of sexual assault cases on Twitter. F*eminism & Psychology, 28*(1), 90–108. https://doi.org/10.1177/0959353517715874

Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012). Predicting Dark Triad personality traits from Twitter usage and a linguist analysis of tweets. In Paper presented at the international conference of machine learning and applications (IMCLA).

Tabachnick, B. G., & Fidell, L. S. (1989). *Using Multivariate Statistics*. New York: Harper & Row, Publishers, Inc.

Urdan,T., (2011). *Statistics in Plain English* (3rd ed.). Taylor and Francis Group.

van de Mortel, T. F. (2008). Faking It: Social Desirability Response Bias in Self-report Research. *Australian Journal of Advanced Nursing 25*(4), 40-48.

van der Bruggen, M., & Grubb, A. (2014). A review of the literature relating to rape victim blaming: An analysis of the impact of observer and victim characteristics on attribution of blame in rape cases. *Aggression and Violent Behavior, 19*(5), 523–531. https://doi.org/10.1016/j.avb.2014.07.008

Von Hohendorff, J., Habigzang, L. F., Koller, S. H. (2017). "A boy, being a victim, nobody really buys that, you know?": Dynamics of sexual violence against boys. *Child Abuse & Neglect, 70,* 53-64. http://dx.doi.org/10.1016/j.chiabu.2017.05.008.

Vonderhaar, R. L., & Carmody, D. C. (2015). There Are No "Innocent Victims": The Influence of Just World Beliefs and Prior Victimization on Rape Myth Acceptance. *Journal of Interpersonal Violence, 30*(10), 1615–1632. https://doi.org/10.1177/0886260514549196

Wei, M., Heppner, P. P., Ku, T., & Liao, K. Y. (2010). Racial discrimination stress, coping, and depressive symptoms among Asian Americans: A moderation analysis. *Asian American Journal of Psychology, 1*(2), 136-150. DOI: 10.1037/a0020157

Wenzel, K., Schindler, S., & Reinhard, M. (2017). General Belief in a Just World Is Positively Associated with Dishonest Behavior. *Frontiers In Psychology*, *8*. doi: 10.3389/fpsyg.2017.01770

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, *9*(11). https://timreview.ca/article/1282