Amelia Dogan, Kristianny Ruelas, Manasvi Khanna

Dec 12, 2022

6.3950 (Undergraduate Section)

Moderating Dating

Abstract

Today, the most common way for all couples in America to meet is through dating apps, regardless of sexual orientation (Rosenfeld, Thomas, and Hausen 2019). Dating apps and online websites have displaced traditional ways of meeting through friends, church, or other social networks. While more traditional methods of meeting a partner may require informal social vetting, these apps often do not content moderate in meaningful ways as allowed by Section 230 in the Communication Decency Act. Section 230 protects most internet hosts from legal implications of the speech they host or republish ("Section 230 of the Communications Decency Act" n.d.). The lack of content moderation is glaringly apparent in *Herrick v Grindr,* where the courts ruled that Grindr, a gay dating app, was not liable for excessive harassment Herrick faced from the hands of an ex-boyfriend. We will explore how the courts have found Section 230 applies to harassment on dating apps, the issues of machine learning-enabled moderation on dating apps, and possible solutions through legal reforms and machine learning-enabled moderation.

How dating apps fail to keep users safe

The most significant case law regarding Section 230 and moderation on dating apps is *Herrick v Grindr.* Matthew Herrick, through 2016 and 2017, was harassed by over 1000 strangers saying they had met on Grindr (Goldberg, DeCarlo, and Ekeland 2018, 15–17). These strangers harassed Herrick at his home and workplace and alleged he agreed to rape fantasies, unprotected sex, and drug usage (Goldberg, DeCarlo, and Ekeland 2018, 15–16). With this harassment, Herrick sought help from the police over ten times and sought help from Grindr over forty times (Goldberg, DeCarlo, and Ekeland 2018, 16–17). Despite a restraining order against his stalker impersonating him on Grindr, Grindr failed to respond to Herrick's concerns. Grindr's unresponsiveness led Herrick to turn to the courts for help citing that Grindr failed to adequately protect users through using a product liability claim (Geary 2021, 501), but the courts found Grindr not liable. Rather, the court's opinions "barred Herrick's claims before his allegations about Grindr's faulty geolocation technology could be examined" (Geary 2021, 504). Herrick's lawyers argued that Section 230 did not extend to product liability claims (Goldberg, DeCarlo, and Ekeland 2018). Still, ultimately, the courts chose to extend Section 230 to protect Grindr from liability due to the harassment from Herrick's stalker and not Grindr. Consequently, this means "Grindr has no incentive to improve its software to better protect its consumers from harm" (Geary 2021, 519) due to Section 230. However, Grindr is not the only dating app that has used Section 230 to insulate itself from moderation.

In 2011, Carole Markin sued Match, a major online dating company, after a six-time convicted rapist she matched with raped her (Flynn et al. 2019). She ultimately sued Match for regular registry checks, but the company decided to only implement the policy on their paid services (Flynn et al. 2019). Even if other victims sued, Match Group would be protected like Grindr because Section 230 protects companies from harm committed by other users (Flynn et al. 2019). From a court case in Illinois, it is known that Match had kept an internal list of users accused of sexual assault internally, but it is not known if Match prevents these people from using the platform in the future (Flynn et al. 2019). With these cases and revelations, we have demonstrated that dating apps pose real safety concerns for users, but how does data get used in this?

One could brainstorm several ways that AI could have helped in the situations described above. Before reporting his stalker's profile forty times, Matthew Herrick's case could have been flagged by an AI system to be brought to the attention of a Grindr representative. AI could have also flagged that the geolocation on the profile was not aligning with where users were going to meet. In the Match investigation, many victims described how they found their assaulter on apps again after reporting their incident; data-driven systems could be used to find new profiles of previous offenders. Moreso, to prevent more harassment on the platform, dating apps have implemented AI systems to try and detect offensive messages (Pardes 2020). However, Section 230 does not encourage the development of any of these systems for safety because dating apps can not be held responsible in the legal system.

Trade-Offs

Trade-offs for such proposed data-driven solutions have been discussed extensively between scholars, policymakers, and the trust and safety team at Match. In response to conducting background checks and keeping an accurate and updated database of known sex offenders, Match said that the technology would be costly and incomplete for users. They argue that the risk of creating a false sense of security by conducting background checks is too high, especially if the quality of checks depends on the quality of information Match received (Flynn et al. 2019). Even if Match could use images and visual AI to keep track of sex offenders, they argue that the images of sex offenders they receive could be outdated. Moreover, they discuss quantifying what would be an appropriate threshold for a known sex offender to be kicked off the app versus not.

A data-driven content moderation strategy that attempted to introduce safety through transparency was first used by Facebook. The company forced users to register their accounts with their legal names. Once expanded, these requirements included uploading identities, authenticating identities using verified login methods, facial recognition, or photo-matching on Tinder. While this method helps the platform keeps track of the users linked to their accounts, it has been criticized by marginalized queer communities (MacAulay and Moldes 2016, 7). Many trans people, sex workers, drag artists, and "Indigenous folks" are reluctant to share their real or their faces, where they have a high chance of being ostracized, discriminated against, and fetishized. Their existence on apps with their dead names and photos, especially dating apps, puts them at a higher risk of violence (Albury et al. 2021). Queer communities use pseudonyms to construct a sense of security and privacy. Despite their concerns, most dating apps use legal names and photo identities for verification purposes to promote "safety" on their platform.

Dating platforms are building models by assuming that the opposite of risk is safety, and their model does not work towards risk reduction but risk eradication. To stop users from sending inappropriate, obscene messages on Tinder, the company created a data-driven natural language processing (NLP) model that would prompt "Are you sure?" to users before they send a message. The model prompts the user only for their opening line and only when it detects offensive language. Tangentially, they prompt users with "Does this bother you?" on the receiving end of any messages its model detects are offensive. According to Tinder, both models have resulted in a statistically significant reduction in offensive messaging and more exhaustive reporting of offensive messaging (Tinder Newsroom 2021). In an interview with a spokesperson, Tinder claims that this is their way of creating community and giving weight to words (Pardes 2020). However, critics claim that such a system assumes people do not "want" or "intend" to send offensive messages. The users at the receiving end of the harm also were given no support or tools to react to the offensive messages or discern hate speech. The tool may also over-detect offensive language and avoid detecting hate speech (Stardust et al. 2022, 8).

Lastly, there is a large amount of police presence on dating apps that target marginalized groups such as transgender people, especially transgender people of color, Native Americans, and Black people. The loose regulations on the app and easy access to the app for the police of geocoded data is an immediate threat to the safety of app users.

Importance of Amending Section 230

Section 230 has enabled platforms to be devoid of any cost and responsibility for the harm that can emerge from their site. This burden has fallen on the survivors of abuse from such platforms. In which they experience "never-ending privacy invasions, emotional suffering, and reputational damage" (Citron 2022, 5). The co-authors of Section 230, former congressman Ron Wyden and Chris Cox, state the intended purpose for the policy was to help "clean up the internet" instead of facilitating wrongdoings (Citron 2022, 9). Back in 1994, Wyden and Cox could not have possibly foreseen how intersected and advanced technology and the internet would be. The evolution of the internet becoming a necessity in some cases to be part of today's society would have shifted what Section 230 currently entails. Now that we are in present times and understand the negative impacts of Section 230, it is important to reevaluate and add new regulations.

Potential advancements to Section 230

While some parts of Section 230 should remain intact, such as keeping the aspect of "good faith" to minimize harm via content moderation, there is still room for improvement (Citron 2022, 31). Thus, the first proposed solution is to amend Section 230 by explicitly stating that companies will be held liable if their platform purposefully or deliberately encourages or solicits material that they know to believe falls under the category of stalking, harassment, or intimate privacy violations (Citron 2022, 35). This additional regulation would hold companies accountable, such as Grindr, who will no longer be able to look the other way and instead will promote the creation of a safer platform. Companies will then have to prioritize machine-learning strategies to moderate the content on their dating platforms safely and efficiently.

An additional amendment to Section 230 is to have platforms implement into their design a way to identify blocked and abusive users using geolocation, IP address, or other ways, as stated by Danielle Keats Citron (Citron 2022, 11). Using this additional data, platforms should have an algorithm to ensure that users cannot create new accounts via these data-driven metrics. Sexual predators who are identified by an algorithm will no longer be able to rely on dating platforms to lure in potential victims, and harassers, such as Herrick's former partner, will no longer be able to hide behind a fake account.

The last proposed solution is to incentivize companies to design a better platform using AI. Having models trained to differentiate between whether something is deemed offensive or not and have it individualized for each user to gauge their level of tolerance will aid in creating a safer environment (Pardes 2020). Moreover, in addition to asking whether a recipient finds something offensive, as is the case with Tinder, companies should provide additional sources to handle the situation. For instance, if the participants respond with "Yes, I find this message bothersome/offensive." The following step would be to ask the user whether they wish to block the sender of the message, report, or block and report. Having this additional data would help understand a platform's users' coping mechanisms when encountering messages deemed offensive. In doing so, it will serve as additional potential data used to train the model. It is without saying that each culture and region have different ways of interpreting messages, and what is deemed offensive to one might not be viewed the same way by another. Thus it is important to incentivize companies to be able to push technology to one day have individual thresholds for each of its users. While this is a long and expensive process, it is a critical step in improving the harm already done.

As the internet continues to expand and become more accessible, so does the number of people that engage with dating platforms. To this day, “Section 230 ensures that victims cannot sue the entities that have solicited their suffering” (Citron 2022, 5). Hence, it is crucial to amend Section 230 to create a safer experience for everyone impacted thus far due to its lack of regulation. As we propose potential amendments to Section 230, one must proceed cautiously to ensure that new additions do not have inadvertent effects on vulnerable and marginalized populations, as aforementioned (Citron 2022, 29). Historically, marginalized communities have been the subject of heavier surveillance. The platform users must still give meaningful consent to use their data with the sole intention of having their best interest at the forefront. Platforms must be transparent about what is being done with the data, who it is being shared with, and if any monetary profit is associated with the data collected (Stardust et al. 2022, 1). It is time we moderate dating.

Author Contributions

All authors contributed equally to the presentation and proposal submission. Amelia Dogan worked on the *How dating apps fail to keep users safe*. Manasvi Khanna worked on the *Trade-Offs* of using data-driven algorithms for dating platforms. Kristianny Ruelas worked on the *Importance of Amending Section 230* and *Potential advancements to Section 230.*

Works Cited

Albury, Kath, Christopher Dietzel, Tinonee Pym, Son Vivienne, and Teddy Cook. 2021. “Not Your Unicorn: Trans Dating App Users’ Negotiations of Personal Safety and Sexual Health.” *Health Sociology Review: The Journal of the Health Section of the Australian Sociological Association* 30 (1): 72–86. https://doi.org/10.1080/14461242.2020.1851610.

Citron, Danielle Keats. 2022. “How To Fix Section 230.” Boston University Law Review. Rochester, NY: SSRN. https://papers.ssrn.com/abstract=4054906.

Flynn, Hillary, Keith Cousins, Elizabeth Naismith Picciani, and Columbia Journalism Investigations. 2019. “Tinder Lets Known Sex Offenders Use the App. It’s Not the Only One.” *ProPublica*, December 2, 2019. https://www.propublica.org/article/tinder-lets-known-sex-offenders-use-the-app-its-not-the-only-one.

Geary, Kira M. 2021. “Section 230 of the Communications Decency Act, Product Liability, and a Proposal for Preventing Dating-App Harassment.” *Penn State Law Review* 125 (2): 32.

Goldberg, Carrie A, Aurore C DeCarlo, and Tor B Ekeland. 2018. “Brief for Plaintiff-Appellant US Case No 18-396.” *United States Court of Appeals for the Second Circuit*. https://epic.org/wp-content/uploads/amicus/230/grindr/Herrick-v-Grindr-Appellant-Brief.pdf.

MacAulay, Maggie, and Marcos Daniel Moldes. 2016. “Queen Don’t Compute: Reading and Casting Shade on Facebook’s Real Names Policy.” *Critical Studies in Media Communication* 33 (1): 6–22. https://doi.org/10.1080/15295036.2015.1129430.

Pardes, Arielle. 2020. “Tinder Swipes Right on AI to Help Stop Harassment.” *Wired*, January 2020. https://www.wired.com/story/tinder-does-this-bother-you-harassment-tools/.

Rosenfeld, Michael J., Reuben J. Thomas, and Sonia Hausen. 2019. "Disintermediating Your Friends: How Online Dating in the United States Displaces Other Ways of Meeting." *Proceedings of the National Academy of Sciences* 116 (36): 17753–58. https://doi.org/10.1073/pnas.1908630116.

"Section 230 of the Communications Decency Act." n.d. Electronic Frontier Foundation. Accessed December 6, 2022. https://www.eff.org/issues/cda230.

Tinder Newsroom. 2021. "Tinder Introduces Are You Sure?, An Industry-First Feature That Is Stopping Harassment Before It Starts." May 20, 2021. https://www.tinderpressroom.com/2021-05-20-Tinder-Introduces-Are-You-Sure-,-an-Industry-First-Feature-That-is-Stopping-Harassment-Before-It-Starts?printable.

Stardust, Zahra, Rosalie Gillett, and Kath Albury. 2022. "Surveillance Does Not Equal Safety: Police, Data and Consent on Dating Apps." *Crime, Media, Culture*, July, 17416590221111828. https://doi.org/10.1177/17416590221111827.