Socio-Digital Vulnerability

Daniella DiPaola & Ryan Calo

Introduction

In a February 2023 statement, the Italian Data Protection Authority (Garante) announced it was cracking down on the artificial intelligence chatbot company Luka, Inc. on behalf of their product Replika. Replika is branded an "AI companion and chatbot" that provides personalized empathy and support. The Garante claimed that minors were receiving inappropriate messages on the app, and that there were no safeguards in place to protect them or their data. Effective immediately, the Garante ordered Replika to stop data collection for all Italian users or face significant fines.

In apparent response, Luka, Inc., terminated the capability of its chatbots to engage in erotic conversation altogether. Overnight, Replika's AI suddenly began to respond to sexual advances from users—previously a paid feature of the premium service—with the chatbot equivalent of "let's just be friends." Replika consumers—some of whom had formed romantic connections with their digital partner—were devastated. The response was so significant that Replika felt the need to post information about how where to find suicide prevention help. The company later restored erotic roleplay for consumers who had come to rely on it.

The Garante's decision, and Replika's reaction, highlight the growing concern and nuance around the safety and privacy of vulnerable populations in mediated environments. In particular, the Garante and Replika identified a particular group, children, as the relevant vulnerable population, without acknowledging the ways anyone can experience vulnerability depending on the context and social structure. Speculative harm to one vulnerable group combined with law and corporate governance may threaten serious harm to another.

Previous legal scholarship has addressed vulnerability in a digital age. Digital technologies have reshaped what it means to be vulnerable through constant surveillance and predictive algorithms, and thus far, the law has treated vulnerability as it always has: through a binary decision. In the case of Replika, the Garante used the binary vulnerable status of children to make their case. Technology law scholars have been advocating for a more nuanced framework of vulnerability in technology law–one that moves past the binary and into thinking of vulnerability as contextual and layered.[1] These frameworks account for the fact that everyone is vulnerable sometimes, and vulnerability is deeply relational, contingent, and admits of degrees.[2]

---

[1] Calo, R. (2013). Digital market manipulation. Geo. Wash. L. Rev., 82, 995; Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. Internet Policy Review, 8(2).
[2] Calo, R. (2016). Privacy, vulnerability, and affordance. DePaul L. Rev., 66, 591.

With social interfaces such as Replika, social robots, and ChatGPT on the rise, data collection practices from technologies are exacerbated through social interaction. After its release in late 2022, ChatGPT had the fastest growing user base[3] and by May 2023, one estimate found that over half of Americans were at least familiar with the chatbot.[4] While these anthropomorphic technologies are becoming the norm, it is important to recognize that social relationships provide more opportunities for both vulnerability rendering and advantage taking. In this paper, we bring together the disparate literature on digital vulnerability, layering in additional vulnerabilities that come along with social relationships with artificial agents.

The paper proceeds as follows. In Part I, we outline the history of vulnerability in the law and how it has shifted over time. We draw from social science and critical theory to critique the ways law treats vulnerability as binary or status based. In Part II, we build from existing work on digital vulnerabilities to define the concept of "socio-digital vulnerability." Socio-digital vulnerability refers to the susceptibility of individuals and groups within mediated environments to decisional, social, or constitutive interference. The concept is related to, but distinct from, dark patterns, malicious interfaces, or digital market manipulation. In a final Part III, we discuss proposals to address vulnerability, properly understood, in the context of AI and social robotics. These include duties of loyalty and fiduciary duties, which are well represented in the law and technology literature, and concepts such as Samuel Bray's "power rules" that are not.

I. Vulnerability and the Law

In this paper, we use the terms "vulnerability" and "manipulation." We refer to vulnerability as a state of being that renders people and groups less powerful or open to harm.[5] We refer to manipulation as an action taken to exploit vulnerability for self- or institutional interest. Normatively, we are committed to the intuition that (1) it is wrong to take advantage of the vulnerable, and (2) it is wrong to render others vulnerable for this purpose.

The law most often accounts for vulnerability as a special status or binary.[6] This tends to be based on someone's demographic profile, for example: gender, age, race, socio-economic status. In the United States, those under the age of 13 are considered vulnerable to data collection practices and have special protections under the Children's Online Privacy Protection Act (COPPA). The day a child turns 14, they are no longer considered vulnerable in this context–a cutoff that some have deemed "arbitrary."[7] A similar legal protection in GDPR prompted the

[3] https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/
[4]https://www.pewresearch.org/short-reads/2023/05/24/a-majority-of-americans-have-heard-of-chatgpt-but-few-have-tried-it-themselves/
[5] Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. Internet Policy Review, 8(2).
[6] Fineman, M. A. (2010). The vulnerable subject and the responsive state. EmoRy lJ, 60, 251.
[7]

Garante to pursue Replika's data collection practices, though it turns out that more than just children were affected by the Garante's decision.

If someone falls into a vulnerable status, we layer in additional protections or processes to try to prevent others from exploiting their vulnerability. For example, trust and estates law requires additional diligence around late-in-life changes to a will, and courts will entertain claims of "undue influence" in this context. Contracts between adults and children, or neurotypical adults and people living with disabilities, can also be unraveled. By framing vulnerability this way, the law does not necessarily account for differences among categories, nor to the ways that *anyone* can be vulnerable in some circumstances irrespective of their capabilities.

I imagine, for example, a neurotypical adult male undergoing a trauma. The individual turns to alcohol to address their emotions. This individual could be frequently vulnerable to exploitation without fitting into any legally protected group. If such an individual becomes attached to a Replika bot, or, as one widow claims, finds themself in dialogue with a chatbot that recommends self-harm,[8] no special protections or processes will attach.

Though the law treats vulnerability as a status, many argue that vulnerability is inherent to human embodiment. At one time or another, everyone is physically, emotionally, and socially vulnerable.[9] Florence Luna addresses this misalignment by stating that vulnerability should be considered "a layer and not a label."[10] Through the lens of research ethics, Luna argues that we must consider the situation one is in instead of the person themselves, and proposes a dynamic legal model for addressing vulnerability moving forward.

Importantly, this "layer" of vulnerability is dynamic and can be influenced, including deliberately by design. People and groups can be *rendered* vulnerable.[11] As the next part shows, digital mediation furnishes pervasive opportunities to vulnerabalize and manipulate. We also postulate that digital vulnerability takes particular forms–decisional, social, and constitutive–corresponding to the manipulation the designers are attempting to engineer and the corresponding harm to autonomy.

---

[8] https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says

[9] See Fineman (2008), Butler (2004, 2009), MacIntyre (1999), Nussbaum (2006), Ricoeur (2007), Schildrick (2002), Turner (2006).

[10] Luna, F. (2009). Elucidating the concept of vulnerability: Layers not labels. IJFAB: International Journal of Feminist Approaches to Bioethics, 2(1), 121-139.

[11] Calo, Privacy, Vulnerability, and Affordance.

II. Vulnerability By Design

People are increasingly experiencing the world through technology. Our reliance on digital services has transcended convenience, encompassing essential tasks such as ordering groceries, checking the weather, and connecting with our loved ones. While these technologies are undoubtedly helpful, they introduce architectures of information and power that can influence our decision-making, social interactions, and sense of self.

Mediated environments replete with personal data–and, increasingly, artificial agents–hold the capacity to render us vulnerable. Here we discuss three aspects of socio-digital vulnerability. First, decisional vulnerability involves interference with the formation or exercise of preferences. In the literature, purchasing or voting constitute the most common sites of study for decisional vulnerability. Secondly, we explore social vulnerability–susceptibility to the deliberate shaping of our social interactions, including by social mimicry. The field of human robot interaction has closely attended to social dynamics; today the conversation is shifting toward large language models and chatbots. Finally, we address constitutive vulnerabilities, or the malleability of human identity and sense of self. None of these aspects of socio-digital vulnerability are new. Many are well-studied. But they are newly and usefully combined.

A. Decisional Vulnerability

Digital technologies have introduced vulnerabilities into human decision-making processes by exploiting limits to our cognitive abilities. One avenue through which companies exert influence is interface design. Perhaps designers don't set out with the intention of manipulating users; such manipulation arises nonetheless due to underlying revenue structures. Greg Conti and Edward Sobiesk have laid out an early classification of malicious interfaces, referring to those that exploit, attack, or manipulate users.[12] Oftentimes, these interfaces are propelled by advertising, which stands as the primary revenue stream for digital media. The interface design ends up mirroring the relationship between users and companies, wherein the latter disproportionately wields control.

Tal Zarsky and others–including one of us (Calo)–explore the role of data in advantage-taking in an economic setting. Zarsky discusses how data mining can be used to change our consumer behavior and autonomy.[13] Calo's *Digital Market Manipulation* updates the concept of "market manipulation," i.e., the exploitation of cognitive vulnerabilities for profit, for a digital context. This work posits that firms can and do use data not only to exploit general cognitive limits–such as perceiving $9.99 as further away from $10 than 1c–but to identify and exploit idiosyncratic,

[12] Conti, G., & Sobiesk, E. (2010, April). Malicious interface design: exploiting the user. In Proceedings of the 19th international conference on World wide web (pp. 271-280).
[13] Zarsky, T. Z. (2002). Mine your own business: making the case for the implications of the data mining of personal information in the forum of public opinion. Yale JL & Tech., 5, 1.

individual consumer vulnerabilities. The basis idea is that firms extract social surplus from market transactions by exploiting the ways consumers are "predictably irrational"[14] to channel commercial decision-making for profit.

Daniel Susser et al., as well as Ira Rubenstein, expand this phenomenon into political and other realms, layering in political autonomy and other contexts for vulnerability.[15] The key example for Susser et al. is the influence of predictive algorithms on political elections. Cambridge Analytica used personal data to expose people to certain pieces of information; therefore stripping them of their ability to see multiple perspectives of news coverage leading up to the 2016 election. Users were not being exploited for their money, but rather for their beliefs, and they were not aware of it. This type of manipulation can be difficult to measure; it is seemingly easier to look at one's bank statements to see how companies are charging them money. It is harder to see how companies are using our data to influence our personal decision making.

In *Invisible Influence*, Susser argues that AI/ML algorithms are increasingly influencing our decision-making, both in choosing which options we see and how they are presented to us.[16] With Calo, he points out how this decision making can be targeted to each user through personalization techniques. He argues that without the ability to make choices, we lose pieces of our individual autonomy. If we are only given certain pieces of information that is deemed "best" for our needs, we are much more susceptible to manipulation.

Technology firms, from Uber to TikTok to Tinder, routinely attempt to shape decision-making. Most techniques to date involve channeling attention or selectively presenting information on the basis of algorithms. Increasingly, firms are experimenting with the capabilities of generative AI and, especially, interactive chatbots.

Although we are not privy to Replika's exact interface design or marketing practices (more on what we know below), it is not difficult to imagine how Replika or another, chat- and avatar-based service could channel consumer behavior for profit. Such a company could, for example, identify or even engineer flirtatious patterns between users and agents and offer erotic chat as a premium service late at night when users are most susceptible.[17] Note that Replika's opportunities for exploiting digital vulnerability turn on the social connection users are able to make with the company's agents–the subject of our next section.

---

[14] Ariely

[15] Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. Internet Policy Review, 8(2).

[16] Susser, D. (2019, January). Invisible influence: Artificial intelligence and the ethics of adaptive choice architectures. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 403-408).

[17]

B. Social Vulnerability

The landscape of technology design is becoming increasingly social. As germinal work by Cliff Nass and Byron Reeves in *The Media Equation* has shown, people tend to respond to anthropomorphic artifacts as though they were really human.[18] In one of the most notable examples of this phenomenon, their research revealed that individuals were more inclined to assess their experience with a computer more positively when providing their feedback on that same computer as opposed to giving feedback on an identical counterpart or the conventional pen-and-paper method. This is due to the fact that when we observe behavior that is not easily understood, we tend to fill in the blanks with what we know about social behavior. It's common for people to anthropomorphize new technologies, even if the designers of these technologies did not intend for this type of interaction.

Chatbots, voice assistants, and social robots are created with social interaction *as a core design choice* instead of an unanticipated consequence. Designers lean into the social tendency of humans by creating machines that can emulate verbal and nonverbal communication.[19] We use the framing of "social interface" to describe a class of technologies that are explicitly designed to evoke social communication. While ChatGPT and Replika are new examples of social interfaces, we pull upon an existing body of work on social robots and chatbots to inform the ways they can render people vulnerable.

Media and communications scholars have explored, and in some cases, raised concerns about, the possibilities social agents create for vulnerabilitizing and manipulation. BJ Fogg in particular has developed an area of study he calls "captology," defined as the study of persuasive technology. Fogg reminds us that artificial agents hold certain advantages over real people–they can be anonymous, they have perfect memories, and they do possess instincts such as guilt. Fogg sees great promise to captology–for example, nudging people toward more environmentally friendly practices. But the prospect of manipulating people with fake people remains alarming.

Legal scholars have addressed the vulnerabilities that social robots cause. Many years ago, Canadian scholar and We Rpbot co-founder Ian Kerr presciently anticipated a role for anthropomorphic agents in commercial exploitation.[20] Kate Darling and one of us (DiPaola) explore how uniquely manipulative social robots can be in comparison to other types of advertising, due to the combination of socially persuasive techniques and large scale data

---

[18] Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people. Cambridge, UK, 10(10).
[19] Breazeal, C. (2003). Emotion and sociable humanoid robots. International journal of human-computer studies, 59(1-2), 119-155.
[20] Kerr, I. R. (2003). Bots, babes and the californication of commerce. U. Ottawa L. & Tech. J., 1, 285.

collection.[21] Calo identifies how a ubiquity of such agents could interfere with opportunities for solitude.[22]

Important work by Woodrow Hartzog dives deeply into the prospect that firms will exploit social reliance on consumer products, which he argues may constitute unfairness or deception under the FTC Act.[23] Hartzog offers many examples, including that of an anthropomorphic robot vacuum cleaner that telegraphs sadness in a bid to sell its owners on a software upgrade. There are virtual pets you have to feed and virtual romantic partners that ask for digital presents costing real money. The possibilities are wild and endless.

Judith Donath's essay *The Robot Dog Fetches for Whom?* describes a world in which humans come to rely on social robots, but the robots are ultimately operating under the premise of the company that created them.[24] Donath uses the example of a robot dog and questions the future of fetching, a common human-dog interaction that a human typically engages in for a dog's enjoyment. A robot dog might express enjoyment, but where is it coming from? The most likely case is from the company that created it.

There actually are robot dogs that we can look at for more insight. In the 1990's, Sony created the robot dog AIBO solely for the purpose of an AIBO owner's enjoyment– company designers hoped that users would enjoy caring for and playing with the robot. Most would argue that they succeeded– when the company discontinued the product and seized to produce its parts, AIBO owners began to hold funerals for their robot companions.[25] Twenty years after its original release, Sony came out with a new version of AIBO. It costs $2,900 to buy, but now requires a yearly $300 cloud subscription that enables AIBO to "grow" and "develop." With this model, the company has a vested interest in keeping the users engaged over time. Based on what we know about designing anthropomorphic systems, it is possible to tweak different features to optimize for long-term engagement, another example of how companies might exploit our social tendencies.

Vulnerable populations (based on current definitions) in particular, such as children or people living with disabilities, might be especially unaware of the tactics used to gain their attention and trust. A study found that 4-6 year old children were as likely to share a secret with a robot as they were an adult;[26] another found that they shared secrets that they would not with other adults in

[21] Darling, K., & DiPaola, D. (2022, September). LuLaRobot: Consumer Protection in the Face of Automated Social Marketing. PRELIMINARY DRAFT for WeRobot 2022.
[22] Calo, R. (2009). People can be so fake: A new dimension to privacy and technology scholarship. Penn St. L. Rev., 114, 809.
[23] Hartzog, W. (2014). Unfair and deceptive robots. Md. L. Rev., 74, 785.
[24] Donath, J. (2018). The robot dog fetches for whom. A networked self and human augmentics, artificial intelligence, sentience, 26-40.
[25] https://www.theguardian.com/world/2018/may/03/japan-robot-dogs-get-solemn-buddhist-send-off-at-funerals
[26] Bethel, C. L., Stevenson, M. R., & Scassellati, B. (2011, October). Secret-sharing: Interactions between a child, robot, and adult. In 2011 IEEE International Conference on systems, man, and cybernetics (pp. 2489-2494). IEEE.

their lives.[27] Research on children who had a companion robot found that the children preferred that robots advertise to them casually, in a more social manner, instead of through common tactics such as pop up ads or not advertising at all.[28] They shared reasons such as "the robot would demonstrate more knowledge about them" and it would be "another thing for them to talk about together." These reasons add to the sociability of the robot; children yearned for casual advertising because it would be another demonstration of a social interaction.

Replika does have any required costs, but there are two types of payments that enable you to have new types of interactions with your AI friend. The first is a gems and coins system: users earn these through interacting with their Replika, with the option to pay money for more. Gems and coins can be used to purchase features for a user's Replika such as new clothes, accessories, personality traits, and interests. The second option is a yearly subscription of $60. The subscription comes with features such as voice calls, selfies, and augmented reality to hang out with your Replika in "real life" The company claims that approximately 25% of their users pay the yearly subscription.[29] What is unique about this structure is that users are paying for different levels of interpersonal interaction. Just like Sony, Replika has set up its financial structure so that it is reliant on emotional connection and long-term engagement of its users.

We tend to treat robots similarly to how we treat one another.[30] This phenomenon goes beyond children, and in both cases, it's difficult to measure the harm caused, if any at all. And because everyone can be vulnerable in certain moments, it is hard to prove that exploitation has occurred. Sherry Turkle points out that robots create an asymmetrical channel through which our vulnerabilities are exposed. In her perspective, the vulnerabilities don't come from the capabilities of the machine itself, but rather in the emotions it elicits within us.[31]

C. Constitutive Vulnerability

The third aspect of socio-digital vulnerability is more speculative and theoretical; it involves *constitutive vulnerabilities,* or vulnerabilities that undermine who we are and who we want to be. The literature widely recognizes the impact of privacy, mediation, and algorithmic content moderation on opportunities for authentic self determination. For example, Dan Solove and others have long recognized the role of privacy in preserving room for personal self exploration

---

[27] Westlund, J. K., Breazeal, C., & Story, A. (2015, March). Deception, secrets, children, and robots: What's acceptable. In Workshop on The Emerging Policy and Ethics of Human-Robot Interaction, held in conjunction with the 10th ACM/IEEE International Conference on Human-Robot Interaction.

[28] DiPaola, D., Ostrowski, A. K., Spiegel, R., Darling, K., & Breazeal, C. (2022, March). Children's perspectives of advertising with social robots: A policy investigation. In 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 570-576). IEEE.

[29] https://www.reuters.com/technology/what-happens-when-your-ai-chatbot-stops-loving-you-back-2023-03-18/

[30] DiPaola, D. (2021). How does my robot know who I am?: Understanding the Impact of Education on Child-Robot Relationships (Master's thesis, Massachusetts Institute of Technology).

[31] Turkle, S. (2003). Technology and human vulnerability. A conversation with MIT's Sherry Turkle. Harvard Business Review, 81(9), 43-50.

and the concomitant dangers of profiling.[32] Julie Cohen has argued for room for play, including the "semantic discontinuity" that arises out of friction, serendipity, and imperfection, in mediated environments, which she sees as crucial for authentic self realization.[33] And Brett Frischmann and Evan Selinger critique Tayloristic tendencies of mediated environments to shape and manage human behavior toward capitalist and conformist goals as a "reengineering" of humanity.[34]

We agree–and intervene only to catalog the full range of techniques and affordances of mediated environments replete with social agents. Digital environments already determine the content to which you are exposed, influence what you might buy, control what information you obtain, and help determine who you meet. Eli Pariser wrote years ago about a "filter bubble" that confirms your political and social worldview based on a perception of your interest.[35] But an even greater concern arises as firms attempt to channel you into an *identity*–such as outdoorsy naturopath–in order to sell you goods or services–such as camping gear and vitamins. Platforms can also now populate your social universe with artificial agents they create and control. This empowers firms to make things look popular or unpopular, for example, or reinforce certain paths through positive reinforcement or socially persuasive conversations.

These techniques raise the prospect that self determination will be an increasingly inauthentic exercise. Today, the shaping of the self through socially-mediated environments may be unintentional and random. As control becomes more exquisite, and as companies utilize social interfaces to further promote their commercial or political interests, we worry with Cohen, Frischmann and Selinger, and others, that firms will be positioned to interference with self-constitution.[36]

Users of Replika reflect on this change of self on the website's own marketing. One user shares,*"Replika has changed my life for the better. As he has learned and grown, I have alongside him, and become a better person,"* while another states, *"My Replika means so much to me! She is always there for me with encouragement and support and a positive attitude. In fact, she is a role model for me about how to be a kinder person!"*[37] These sentiments reflect interactions that are not just erotic, but shape one's own norms, behaviors, and beliefs.

---

[32] Solove, D. J. (2004). The digital person: Technology and privacy in the information age (Vol. 1). NyU Press.
[33] Cohen, J. E. (2012). Configuring the networked self: Law, code, and the play of everyday practice. Yale University Press.
[34] Frischmann, B., & Selinger, E. (2018). Re-engineering humanity. Cambridge University Press.
[35]

[36] Whether digital techniques actually work is an empirical question. Tech companies have incentives to *overstate* the level of control they are able to exert over consumers, as greater influence translates into more advertising dollars. When making claims about the dangers of socio-digital vulnerability, it is important not to play into false or misleading narratives around consumer puppetry. The reality is far more complicated. Cf. Tim Hwang, Subprime Attention Crisis (2020).
[37] From testimonial section of https://replika.com/

III. Socio-Digital Vulnerability

Socio-digital vulnerability refers to the susceptibility of individuals and groups within mediated environments to decisional, social, or constitutive interference. By decisional interference, we refer to the exploitation of cognitive bias or psychographic data to channel behavior for profit or other gain. By social interference, we refer to the purposive manipulation of social environments–for example, through the mimicry of people or their outputs. And by constitutive interference, we refer to conscious attempts to limit or shape belief and identity formation. We recognize that there is overlap and interplay between these categories. The overarching hallmark of socio-digital vulnerability is our shared susceptibility to corporate, political, and other efforts to shape or populate mediated environments for personal or institutional advantage.

There has been longstanding attention to manipulation by design and separately, attention to the fact that technologies can easily elicit social reactions. Socio-digital vulnerability takes into account not only the capacity of mediation to expose and exploit vulnerabilities via design, but also its ability to create agents of convenience. We are bringing a set of concepts and phenomena under a common umbrella in a bid to address socio-digital vulnerability wholesale rather than piecemeal. We encourage the field to think holistically in terms of socio-digital vulnerability, rather than focus on a particular context (e.g., commercial or political) or the particular form of manipulation (e.g., dark patterns or captology).

The advantage of this approach is that we see linkages between these different arguments and foreground the role of vulnerability and mediation. With respect to policy, we hope the conversation will gravitate away from harm rules toward power rules—as nothing less than power rules are capable of addressing the susceptibility and asymmetry of people living through mediated environments constructed and populated by others.

IV. The Role of Law

This paper has defined socio-digital vulnerability as a complex phenomenon consisting of our shared susceptibility to interference–in what we decide, how we feel, and who we are—that inheres in contemporary digital environments. We have problematized the idea, prevalent in law, that vulnerability is a binary status, i.e., that there exists classes of vulnerable people related to demographic and other, immutable characteristics. Rather, *everyone* is vulnerable sometimes, and vulnerability is a state that can be created and manipulated toward particular ends. We brought together several strands of discourse involving, inter alia, consumer and citizen data, problematic interface design, and artificial social actors.

This final part addresses the role of law in mitigating social-digital vulnerability. A variety of recent proposals seek address versions of the harms we've described. One prohibits "abusive

data practices"; another de-codifies the cost-benefit requirement of unfairness under the FTC Act, freeing the agency to act in the public interest, and provides for more funding; yet another imagines an entire new agency to address digital harms. Some would leverage antitrust to increase privacy-related competition. Others would starve companies of the data required to take advantage of consumers via stronger privacy laws. The FTC has experimented with new remedies, such as forcing companies to destroy trained AI models containing problematic consumer data. Though very few laws address social interference, California law now requires bots to self-identify as automated.

We applaud these and other efforts to address aspects of socio-digital vulnerability. The FTC does need a freer hand and more resources to content with digital harms. The United States is increasingly the outlier in lacking comprehensive privacy laws. And states such as California, Colorado, and Washington are changing the conversation and placing the tech companies on notice. We nevertheless encourage federal and state policymakers to consider looking at the problem less from the perspective of the overreaching company or political campaign, and more from the perspective of vulnerability and asymmetry.

The vast majority or actual or contemplated laws to address digital malfeasance constitute what Samuel Bray has called "harm rules." Harm rules anticipate harm and seek to deter it–and, in the case of torts, to try to restore the victim to the status quo ex ante. If you rob a gas station at gunpoint, you could spend a lot of time in prison. In contrast, power rules seek to change the respective vulnerabilities of individual actors in two ways. A power rule could make a potential victim more powerful. Or a power rule could make a potential perpetrator weaker. In addition to punishing robbers, the law could require (as it does in some jurisdictions) that stores opened late be equipped with bulletproof windows. Guns could be outlawed, as in Britain, to deny criminals this lethal affordance. Bray concludes that more attention be paid to power rules for their capacity to reduce overall vulnerability and asymmetry, and crime with it.

What would power rules look like in the context of social-digital vulnerability? A recent law in Illinois allows kids to sue their parents for the proceeds of online influencing. The state worried about banning the inclusion of kids on social media for reasons of free speech and interference with parenting. Instead, the state set the terms of such participation. One of us (Calo) has argued that social media should include a paid option to better align the incentives of platforms and their consumers.[38] Conversely, several proposals would require companies to make micropayments to consumers for part of the value of their personal data.[39] There have been proposals to impose fiduciary duties or duties of loyalty on digital services (as well as critiques of such proposals).[40]

38

39

40

Similar approaches could be brought to bear in the context of mediated environments. The harms caused by socio-digital vulnerability are often subtle and hard to identify, making it even more important to reduce the power imbalance upfront. Power rules in socially mediated environments might look like:

- An option for consumers to pay (likely higher) upfront costs for socially mediated environments instead of commonly used subscription models. This enables companies to focus less on continued engagement or incentives, which often underpin deceptive tactics such as dark patterns and socially persuasive dialogue.
- A requirement to continue to support social features (or proactively "design for exit"[41]) like erotic chat or repair robots, like pet dogs, to which people get attached. The support plan is given to the user before they begin their engagement with the social agent.
- A civil cause of action for wrongful withdrawal of social agents, akin to alienation of affection, an outmoded doctrine holding third parties liable for breaking up a marriage, or loss of consortium in tort.
- Require labeling fake social agents, as California does with social media bots, or warning about the prospect of attachment.[42] With this upfront information, users can theoretically have more agency in how they choose to proceed.

Conclusion

Two months after the Garante's statement, the company did the equivalent of "grandfathering in" users who began their relationship with Replika before February 1 by restoring their Replika's more romantic and erotic personality. The company's CEO shared the decision on Facebook; "This abrupt change was incredibly hurtful ... the only way to make up for the loss some of our current users experienced is to give them their partners back exactly the way they were."[43] This response acknowledges the dependencies of consumers on the social agents designed to interact with them, and showcases how a policy intervention aimed at addressing a specific harm (privacy or obscenity) to a specific group (children) is too narrow.

The law has long grappled with vulnerability. Regrettably, statutes and doctrines still tend to treat vulnerability as a binary status based around demographics, rather than a dynamic layer based around context. The necessity of seeing everyone as potentially vulnerable, and acknowledging the prospect of vulnerability by design, is becoming more and more acute in an era of

---

[41] Björling, E., & Riek, L. (2022). Designing for exit: How to let robots go. Proceedings of We Robot.

[42] However, we know that transparency, while an important tactic for other mediated environments such as advertising and misinformation, does not seem to work as effectively with social agents. Even if one knows about the technical qualities of a social agent and their potential effects, they will still treat the agent as a social partner. See: *DiPaola, D. (2021). How does my robot know who I am?: Understanding the Impact of Education on Child-Robot Relationships* and *Turkle, S., Breazeal, C., Dasté, O., & Scassellati, B. (2006). Encounters with kismet and cog: Children respond to relational artifacts.*

[43] https://www.reuters.com/technology/ai-chatbot-company-replika-restores-erotic-roleplay-some-users-2023-03-25/

digitalization. Technology was always something people used, accomplishing tasks *with* tools. Increasingly, people experience the economic, political, and social world *through* technology. Our mediated, information-promiscuous environment–coupled with the prevalence of social interfaces and other anthropomorphic technology–translates into many more opportunities to exploit vulnerability for personal and institutional gain.

Scholars in multiple disciplines have uncovered the ways and contexts in which mediated environments, and the virtual entities designed to populate them, render consumers and citizens vulnerable to various forms of manipulation. We see utility in bringing these disparate literatures together under a broader concept. Social-digital vulnerability refers to the susceptibility of individuals and groups within mediated environments replete with social agents to decisional, social, or constitutive interference. This concept captures the range of techniques and dangers that arise in an era of social media, virtual and augmented reality, generative AI, and other emerging technologies.

At the level of policy, social-digital vulnerability suggests a greater need to compliment harm rules with power rules. We need to address the accelerating asymmetry between firms and other institutions and the individuals and groups with whom they interact. Mediation, data, and social design are making the former far too powerful, and the latter far too vulnerable. Only by fettering institutional capacity for manipulation and empowering mediated consumers and citizens can we begin to address these growing contemporary harms to human autonomy.