Minds and Machines (2026) 36:28
https://doi.org/10.1007/s11023-026-09784-x

Empathy in the Digital Age: Implications for Content Moderation and Platform Governance

Bhanuraj Kashyap[1] · **Paul Formosa**[1]

Received: 10 March 2025 / Accepted: 13 May 2026 / Published online: 23 May 2026
© The Author(s) 2026

Abstract

Concerns about the toxicity of the internet and social media are widespread. Many reasons have been offered for why social media increases online toxicity, from bad algorithms to the rise of online trolling, but perhaps the most prominent explanation has been the negative effects of social media on empathy. In this paper, we offer a framework of digital empathy by first outlining a *descriptive* account of digital empathy, encompassing both affective and cognitive components. By "digital empathy", we mean empathy mediated through digital environments. We also explore the *normative impacts* of platforms on digital empathy and offer four main categories of moral assessment: (1) morally problematic one-off failures of digital empathy; (2) aggregate harms of failures of digital empathy; (3) impacts of recommendation systems on digital empathy by amplifying, obscuring, framing, and rushing targets of empathy; and (4) moral deskilling of empathy due to widespread use of social media. We also critically explore harm-based justifications for moderating content to explicitly govern empathetic relationships between users of a platform. We demonstrate how discussions of digital empathy can inform and enhance broader debates about the ethics of online content moderation, thereby further illuminating the moral, social, and political implications of social media's role in facilitating user interactions.

Keywords Social media · Digital empathy · Platform governance · content moderation · Recommendation systems

✉ Bhanuraj Kashyap
bhanuraj.kashyap@mq.edu.au

Paul Formosa
paul.formosa@mq.edu.au

[1] Department of Philosophy, Macquarie University, Sydney, Australia

1 Introduction

Concerns about the toxicity of the internet and social media are widespread. For example, in an article in *Policy Options*, Maite Taboada (2021) laments that the internet has provided the opportunity for everyone to become a bad actor and "ordinarily toxic". In *The Atlantic*, Jonathan Haidt (2022) blames social media for internet toxicity in the US and states that social media has "unwittingly dissolved the mortar of trust, belief in institutions, and shared stories that had held a large and diverse secular democracy together". We understand toxicity here to refer to hostile, abusive, and degrading features of digital interactions that cause harm to participants or make online environments less safe for constructive engagement (see Taboada, 2021). Many reasons have been offered for social media driving toxicity, from bad algorithms to the rise of online trolling. However, one of the most prominent explanations for the rise of internet toxicity has been the negative effects of social media on empathy, where a reduction in our capacity to empathise is both a cause of toxic behaviour online and an effect of such toxicity. For example, an article from *The Washington Post* suggests that most of our social media problems stem from an empathy gap (Rubin, 2024), while another article urges users to address the empathy deficit caused by social media use (Hall & Leary, 2020). Talk of the demise of empathy is seemingly everywhere, and social media is portrayed as the key suspect for this demise.

While psychologists and social scientists have focused extensively on empathy in the digital world (see, e.g., Konrath et al., 2023), barring a few exceptions discussed below, philosophers have done far less to contribute to these important debates around digital empathy, even though empathy itself is a much-theorised concept in moral philosophy (e.g., Smith, 1976; Maibom, 2007; Slote, 2010; Formosa, 2017). In this paper, we fill this gap by developing a philosophically robust account of digital empathy. By "digital empathy", we mean empathy mediated through digital environments. Given the importance of social media, we limit our discussion of digital environments to social media platforms (Srnicek, 2017), such as Facebook, YouTube, and TikTok. The existing philosophical literature on digital empathy (e.g., Torbjørnsen & Hipolito, 2024; Mossner & Walter, 2024; and Ventham, 2024) tends to focus on *affective* empathy (i.e., feeling *with* or *for* someone). Our account goes beyond this approach by including both the affective and the *cognitive* (i.e., putting yourself in someone else's shoes) components of digital empathy. We also combine discussions of the significance of social media's design (e.g., like/dislike buttons or the length of characters of a Tweet) with the literature on recommendation systems, which allows us to subsume our discussion of digital empathy under larger conversations about content moderation on social media platforms. Finally, given concerns about the moral value of empathy (see, e.g., Slote, 2010; Bailey, 2022), our account develops a pro tanto justification of the moral value of digital empathy grounded in its ability to express *moral connectedness* between agents.

The outline of our paper is as follows. In Sect. 2, we engage with the existing literature on empathy and content moderation. By drawing on this literature, we develop in Sect. 3 our descriptive account of digital empathy, before offering a moral assessment of digital empathy in Sect. 4. Finally, Sect. 5 situates our account of digital

empathy within the broader content moderation literature, enabling us to examine the moral justification for moderating digital empathy on social media platforms.

The central claim of the paper is that digital empathy is a philosophically important but undertheorized phenomenon, and that social media platforms, through their design and recommendation systems, generate four distinctive categories of moral harm to digital empathy that have significant implications for content moderation. Our paper thereby makes three key contributions to the literature. First, we propose a conceptually clear framework of digital empathy and offer a moral diagnosis of the state of digital empathy, and by extension, non-digital empathy in our society. Second, we identify and delineate four key moral issues with digital empathy that arise due to social media (see Sect. 4). Third, we situate our discussion within the larger project of the philosophical issues that emerge from social media content moderation (see, e.g., Lazar, 2023). By doing so, we urge philosophers to not only think about the flow and distribution of information through social networks on the internet (e.g., filter bubbles, echo chambers and so on), but also to consider the moral issues that arise due to the curation of relationships that obtain between users on social media.

2 From Empathy to Recommendation Algorithms

2.1 Empathy

Empathy is a multifaceted phenomenon that helps us to know what other people think and feel, understand their point of view, and participate in their sorrows and joys (Maibom, 2014; Cuff et al., 2016). Empathy is often analysed along two dimensions or components, cognitive empathy and affective empathy, that can co-occur or exist independently. First, *cognitive* empathy refers to the ability to understand another's perspective by putting yourself in their shoes. For example, when we imagine ourselves from the viewpoint of someone with whom we disagree, we exercise cognitive empathy toward them. Second, *affective* empathy refers to experiencing emotions *with* or *for* another person. For example, we experience affective empathetic anger by mirroring our friend's anger that they experience at their own suffering, and we experience affective empathetic joy by mirroring our friend's joy at a piece of good news they have received.

Further, empathy is often contrasted with nearby concepts, such as sympathy and compassion, which are seen as feeling *for* someone else (Darwall, 1998). In contrast, empathy is seen as feeling *the same thing* as others.[1] In line with much of the non-philosophical literature and common linguistic use (see e.g., Denham, 2017, 230–231), we shall understand affective empathy broadly here to include both sympathy and compassion, and thus include cases of both feeling *with* (e.g., we both feel anger

[1] Emotional contagion, personal distress, and pity are some of the other nearby concepts (see Maibom, 2014, and Vallor, 2012, for discussions on the various distinctions between empathy and these nearby concepts) that get discussed in the philosophical literature on empathy. We set these aside here because, unlike sympathy and compassion, they do not directly deal with the other-directed perspective taking or concern that is central to our account of digital empathy and moral connectedness that we give below.

at your mistreatment) and *for* others (e.g., I feel angry at the mistreatment you suffer which causes you sorrow).

It is worth stressing that many instances of empathy invoke both components: for example, when we cognitively imagine ourselves in another's shoes, we may also feel anger at the injustices they endure once we have properly grasped their point of view. However, these two components can also come apart, since sometimes we can feel what others feel without first imagining ourselves into their viewpoint, and we can sometimes imagine ourselves into their viewpoint but feel nothing. The concept of empathy—or its close correlates, such as sympathy—has a rich history in philosophy. Philosophers have employed the concept of empathy to address the problem of other minds (Husserl, 1970; also see Goldman's simulation theory, 2006), to explain our altruistic motivations (Ventham, 2023), to establish the grounds of morality (Hume, 1975; and Smith, 1976), as a capacity to be cultivated as part of an imperfect duty (Kant, 1996, 6:456−62; cf. Formosa, 2017), and as a key input into moral judgement (Slote, 2010; Kauppinen, 2017). However, while empathy is widely regarded as fostering prosocial judgements, limiting harm to others, and promoting virtues such as friendship, loyalty, care, and benevolence (Denham, 2017), there are also strong criticisms of attempts to give empathy a more foundational role in morality. For example, Prinz (2011a, b) argues against empathy as the ground of moral judgment by noting that we make moral judgments about crimes that have no victims to empathise with and consider some acts as impermissible despite empathising with the perpetrators of those crimes. For these and other reasons, both sentimentalists, such as Prinz (2011a) who offers an account where moral judgements are best explained by emotions such as anger and moral indignation without requiring a prior empathetic response, and rationalists of various stripes (see Kennett, 2017; Maibom, 2014) have been disinclined to view empathy as *foundational* for moral judgement. Furthermore, empirical research indicates that biases and identity markers influence who we empathise with. For example, we are more likely to empathise with individuals from our own racial group than a stranger from another racial group (Avenanti, Sirigu, & Aglioti, 2010). Empathy's partiality (Kauppinen, 2014) can mislead and prevent us from reaching robust moral judgements about those we are less likely (or able) to empathise with.

Despite the problems with placing too much moral weight on empathy noted above, a strong case can still be made that empathy is morally valuable and plays an important, if not foundational, role in a moral and virtuous life. To make our case, we focus on two features of empathy. First, empathy, especially cognitive empathy, has *epistemic* value by enhancing our understanding of the perspectives and experiences of others (Schramme, 2024). When we empathise with others, we apprehend their emotionally charged perspective, which can help us to see them as "creatures whose emotions, however irrational they may be, nevertheless make perfect human sense" (Bailey, 2022, p. 61). This understanding, which Bailey (2022) terms "humane understanding", has moral implications. We understand "humane understanding" here to be an *epistemic achievement* gained through the cognitive recognition that the other's experiences and point of view are intelligible from their perspective. For example, imagine empathising with a distressed homeless man. In doing so, you see the world from his perspective, marked by the evaluative features characterised by emotions such as anger and hopeless-

ness. If the homeless man is in a fit of rage and screams at passersby, your empathy might help you to see his anger as *intelligible*, from his point of view, without necessarily endorsing his behaviour. When we empathise, we recognise others as having intelligible subjective experiences, even if we disagree with their judgments or actions. In doing so, their experiences are rendered coherent, and we end up acknowledging them as epistemic agents, whose perspectives and testimonies about their own lives merit credibility. This helps us take them seriously as genuine sources of knowledge about their own experiences (Fricker, 2007). Second, the affective component of empathy amplifies our connections with others by alerting us to morally salient features that we might otherwise overlook (Formosa, 2017, 196–198; Bailey, 2022). For example, empathy can alert you in real-time to the anger and distress another person is suffering and help you to understand the sense of injustice and hopelessness that may be driving their behaviour. This can alert you to moral dimensions of the situation that you might have otherwise been unaware of, overlooked, or not noticed in real-time.

However, the mere activation of cognitive or affective empathy does not necessarily secure the normative goods associated with this capacity. Empathy functions as a success term: one must meet certain criteria to be empathetic *at all* (as opposed to, e.g., ignoring others), and a more stringent criterion needs to be met to count as empathising *well* (see Bailey, 2022, p. 57). More stringently, empathising *well* involves both accurate humane understanding and modulating an appropriate affective response. Empathising *well* is thus a normative success condition, and we can fail to meet this demanding condition in multiple ways. For example, we can be bad at reading others' body language, misread their emotions, struggle to imagine ourselves in their situation, fail to see how their responses may be intelligible from their point of view, miss morally relevant features of their perspectives and experiences, or fail to properly modulate our affective responses.

Two crucial features of empathy must be stressed here. First, empathy is a skill that requires practice and learning to engage with the perspectives of others (Vallor, 2016). For example, consider a scenario where a child mocks someone in a wheelchair. To foster the child's empathy, we could guide them through experiential learning. This could involve conversations with other children who use wheelchairs, reading books about what it is like to be a child who needs to use a wheelchair, or even firsthand attempts at navigating the world in a wheelchair. Through these experiences, the child can develop a better understanding of life from a wheelchair user's viewpoint. This process not only cultivates the child's empathy but also encourages them to empathise well by accurately grasping the lived experience of another and becoming attuned to the morally significant aspects of their situation. As a result, the child may develop genuine care and concern for wheelchair users, and they may also come to see their past mockery as shameful.[2] By learning to empathise well in this

[2] In this example, we are referring to wheelchair users as a category of potential targets of empathy. For the child to empathise well with a specific wheelchair user, they would also have to accurately empathise with this specific wheelchair user's emotions at a particular time.

manner, the child is developing skills in relating and connecting to others in morally valuable ways.

The second crucial aspect of empathising well is the ability to self-regulate our empathetic responses (Kauppinen, 2014). This self-regulation involves adjusting our empathy levels, both down-regulating or diminishing our empathetic responses toward those we already care about, and up-regulating or amplifying our empathetic responses toward strangers or those we find hard to identify with. Such calibrations help to address concerns about empathy's partiality, which can manifest as either excessive or deficient empathetic responses. For example, down-regulating your empathetic responses towards fellow citizens whom you already care about (e.g., by making their concerns temporarily less salient to you) may help you to up-regulate your empathetic responses towards incoming refugees whom you know little about (e.g., by making their plight and experiences more salient to you), and can give you a clearer moral understanding of a situation. However, this process of up-regulating empathy, particularly towards unfamiliar groups, carries risks. For example, insufficient or misleading information about the individual experiences and perspectives of refugees might lead to misplaced emotional responses, such as feeling anger at refugees for not remaining in their homeland to fight against tyranny or endure extreme poverty, rather than leading you to empathise with the difficulty of their situation and feel sorrow at their plight. While self-regulation of empathy is thus challenging, it is nonetheless an essential component of learning to empathise well.

When we successfully empathise well by achieving accurate humane understanding and appropriately regulating our affective responses, we unlock a distinct *relational achievement* that we term "moral connectedness". To understand this relational achievement, we must differentiate it from a baseline moral commitment to others. We owe all other persons a baseline moral commitment to regard them as having worth, dignity, and basic moral rights, independent of our capacity or willingness to empathise with them. When, for example, we empathise well by grasping the intelligibility of another's experience of suffering and feel appropriately modulated affective empathic pain for their plight, we show in a very direct way that the other person's suffering really matters to us. The moral connectedness displayed by this relational achievement actively instantiates and *enriches* our baseline moral commitment to regard others as the bearers of dignity and creates a thicker relational connection with them.

The distinction between a baseline moral commitment and the relational achievement of moral connectedness is particularly evident in extreme cases where we have strong reasons *not* to empathise with others, such as fascists or white supremacists. In such cases, while we still owe such individuals a baseline moral commitment (such as respecting their basic human rights), we have strong moral reasons to actively withhold (or not seek) the thicker relational state of moral connectedness with them. In such cases, rather than seeking to empathise well with them, it might instead be more morally appropriate to express moral outrage and anger toward them (Hirji, 2022). While such cases show us that empathising well with others may not always be morally valuable *all things considered*, we can still hold that empathising well is at least *pro tanto* morally valuable since in many cases seeking such a thick connected relational state with others is a morally beneficial way of instantiating an enriched

moral commitment to their worth as persons. Moving forward, in Sect. 4, we will use this understanding for our critical analysis and moral assessment of digital empathy. But first, we need to examine the digital context of digital empathy by exploring the recommendation algorithms that drive our experiences of social media.

2.2 Recommendation Algorithms

To understand the impact of social media on digital empathy, we must first comprehend how social media operates. Social media platforms, such as Facebook, Instagram, and YouTube, bring together people from all walks of life and provide their users with a digital space designed to offer various opportunities for sharing content and interacting with one another. Social media is the new public sphere (Cohen & Fung, 2021); it shapes our speech, communicative norms, and ways of relating to one another (Lazar, 2023). Recommendation algorithms, in turn, are "the algorithms driving social media" (Narayanan, 2023, p. 5) and other systems.[3] Recommendation algorithms work by personalisation: they recommend content to a user based on a user's history and the type of content other *similar users* are engaging with or are likely to engage with. By directing a user's attention to recommended content, these algorithms encourage users to spend more time on a given platform, thereby creating opportunities to commercialise their attention and data (Hindman, 2018). Platforms then generate income by targeting users with personalised advertisements, or by creating elaborate user profiles that can be sold to data brokers, or both (Benn & Lazar, 2022).

It is worth emphasising that recommendation systems serve a practical purpose by assisting users in navigating the *infoglut*, which is understood as "an era of information overload" (Andrejevic, 2014, p. 1). Social media platforms are teeming with a near infinite amount of content. Previously, social media relied more on subscription (e.g. accessing content published by the users one follows) and network (e.g., accessing content published by members of the user's online network) models of content distribution, but now a significant amount of engagement is the result of recommendation algorithms (Narayanan, 2023). These algorithms are a response to the infoglut: users' attention is directed to niche personalised content that they might otherwise miss, thereby encouraging users to spend more time on a platform.

Besides the *filtering* of content, another key function of recommendation systems is to *amplify* engaging content. When a piece of content is amplified, it reaches a significantly larger number of users than it would otherwise. Sometimes, content can also go *viral*. While it might be challenging to draw a strict conceptual distinction between amplification and virality, content that goes viral often gains incredible online visibility and popularity. Virality promotes the development and encouragement of various online trends, such as the Ice Bucket Challenge, and fuels new and emerging cultural phenomena, including video responses to new music and movies on YouTube. Recommendation systems are also essential for online creators and

[3] Outside of social media, recommendation algorithms are also deployed across search engines and streaming services such as Netflix—see Narayanan (2023) for a comprehensive guide to recommendation systems.

influencers to distribute their content widely, increase their subscribers, and generate income from the proliferation of their content.

In opposition to amplification, recommendation systems are also involved in *demoting* content. When a piece of content is deemed problematic by a social media platform, it may be downranked, resulting in it being shown to fewer users (Gillespie, 2022). Note that demotion is best construed as a content moderation technique (Kashyap & Formosa, 2023). However, since demotion overlaps with recommendation systems, the lines between content moderation and recommendation systems often blur. Recommendation systems, unsurprisingly, have become the cornerstone of platform analysis in recent years (Keller, 2021; Lazar, 2023), with many commentators arguing that most ills associated with social media are, in fact, attributable to the recommendation systems employed by those platforms. By fixing these systems, it is argued, we can alleviate social media ills (see, e.g., Howard, 2024).

We are, however, sceptical that we can fix all the ills of social media by changing recommendation systems. A minor adjustment to a recommendation algorithm can lead to unforeseen consequences that may be more detrimental than the issues the designers sought to address (Narayanan, 2023). Moreover, viewpoint-based algorithmic demotion of speech may raise moral and legal concerns regarding infringement on users' right to free speech (Keller, 2021). Nonetheless, despite these cautions, we are on board with the larger project of paying critical attention to the contribution of recommendation systems in structuring social media experiences. We extend this literature in a novel direction by analysing the role of recommendation systems in shaping digital empathy.

3 Descriptive Account of Digital Empathy

As noted above, scholarly theorising about digital empathy has largely focused on digital *affective* empathy (see Torbjørnsen & Hipolito, 2024; Mossner & Walter, 2024; and Ventham, 2024). By differentiating between cognitive and affective empathy in the descriptive account that follows, we aim to provide a more comprehensive understanding of digital empathy. But before outlining that account, we need to make a general point about the empirical evidence concerning the relationship between empathy and social media.[4] Numerous empirical studies conducted in the US have reported evidence that social media use is associated with lower empathy (Guan et al., 2019).[5] As we have become more digitally connected over the last two decades, some of our connective abilities have subsequently declined, at least as reported in the US. This is known as the "empathy paradox": we are now more digitally connected, but this has coincided with reported declines in individual empathy (see Konrath, 2013). Furthermore, this

[4] Note, many empirical studies do not distinguish between empathy and digital empathy. We have chosen to continue with this convention when summarising this empirical literature in the present section.

[5] Of course, it is not only empathy that has been impacted by this change, but it is the only factor we focus on here. For a broader discussion of anxiety and other potential impacts, see e.g., Vuorre and Przybylski (2023); Haidt (2024).

drop in overall levels of empathy overlaps with a rise in narcissistic traits in populations, especially young adults who are inclined to spend more time on social media (Konrath, 2013, 208–210).

While many of these studies attempt to show a correlation between a decline in empathy and the advent of social media, they do not conclusively demonstrate that social media is the *cause* of that decline. One of the empirical methods used to assess this decline in empathy measures generational changes in empathy since the 1970s (Konrath, 2013). This means that the rise of social media may be *merely coincidental* with the reported decline in empathy, leaving open the possibility that other factors have contributed to and caused this decline. Moreover, contrary evidence exists (see Vossen & Volkenburg, 2016; Konrath et al., 2023), and the concept of empathy is often ambiguously and inconsistently deployed across studies on empathy (Cuff et al., 2016). Since the empirical evidence is in flux, we believe that it does not (yet) support all things considered assessment of social media's impact on empathy in general and digital empathy in particular (although the evidence may be strengthening—see Haidt, 2024). Given that the empirical evidence we have presented remains contested, we do not place too much argumentative weight on it. To proceed, our argument only requires the more modest claim that the status of empathy in online spaces is not as good as it could be, whether or not it has been significantly worsened by increased social media use.

3.1 Digital Cognitive Empathy

Digital cognitive empathy refers to digitally mediated cognitive empathy: the ability to imagine oneself in another's perspective via digital means. For example, someone who has never owned a dog, but watches videos on Instagram of dog owners playing with their dogs, may, as a result, become able to imagine themselves owning a dog. This act of imagining herself in the shoes of a dog owner, mediated by a digital platform, to make sense of the relationship dog owners have with their dogs, is an example of digital cognitive empathy. The more information you have about such relationships, the easier it seems, *prima facie*, to imagine yourself in the shoes of a dog owner.[6] To deepen your empathetic engagement, you could also fill in more details in your imagination to match the experiences of a dog owner: feeding your dog, recording him playing with strangers, and teaching him to sit or stay. When done well, you might also experience affective empathetic emotions of warmth and affection towards other pet owners as a result of your digital cognitive empathy, which is mediated by watching their videos on social media platforms.

By giving us more information about these diverse experiences, social media has the potential to help us get better at empathetically imagining ourselves into the lives of others. This includes multiple axes of diversity. For example, some of these social media users are geographically and culturally diverse (e.g., users in Brazil and Uganda debating about football), some are ideologically and politically diverse

[6] Please note that we return to the question of information overload in digital spaces in Sect. 4.3 to further develop our analysis.

(e.g., conservatives interacting with progressives), some have diverse socioeconomic statuses (e.g., a wealthy CEO and a poor farmer engaging over a shared interest in kombucha fermentation), and still others are diverse along social identity dimensions (e.g., young queer people commenting on posts shared by older feminists). Moreover, social media accommodates and encourages interactions between its users, including those from diverse backgrounds. Users tweet, post, and share photographs, write commentaries, edit blogs, and create and share videos of themselves engaging in activities and conversations with others. Sharing content about one's life might lead other users to respond with content about their own lives and diverse experiences.

On most platforms, the content and thus the users that you engage with are algorithmically personalised and selected, driven by the platform's recommendation systems. Users are recommended content that algorithms predict will lead to some form of optimisation, whether that be watch/listen time or advertisement and subscription revenue. This process unfolds not only from recommending tokens of content to a user, but also from demoting and, at times, censoring other tokens of content. For example, your recommendation algorithm on YouTube might uprank cat videos, downrank videos about exercise, and remove war videos from circulation altogether. This has also led to the filter bubble concern, where users come to have their own "unique universe of information" (Pariser, 2011, p. 10), thereby disconnecting them from alternative sources of information and from evidence that contradicts their beliefs and values.[7] If people *predominantly* acquire and access information from those within their "bubble", this will not only raise epistemic concerns (see Nguyen, 2020; Ranalli & Malcolm, 2023)[8] but also have implications for empathy, as discussed below. Although social media content moderation is dynamic, unpredictable, and suffers from explainability issues (Gillespie, 2018), one thing is clear: *which* perspectives and users you might come to empathetically engage with on a platform are tailored for you by (often) opaque recommendation systems.

3.2 Digital Affective Empathy

Digital affective empathy is affective empathy mediated by digital means. For example, you experience digital affective empathetic sadness when you watch a video on YouTube and feel sad for the person in the video who is detailing a difficult mental health issue. However, digitally mediated encounters tend to be severely limited in the amount of information they convey compared to their offline equivalents. As Mossner & Walter (2024, 5) note, "digital spaces have (as of yet) neither the same *depth* nor *diversity* than the non-digital world", and digi-

[7] Many contend that the empirical evidence for filter bubbles is inconclusive, suggesting that we should be sceptical, pending further evidence, about filter bubbles and their "supposed" effects on users (see Arguedas et al., 2022; Barrett, 2022).

[8] The epistemic concerns bleed into our concern about failures of empathy: when a hearer believes she is justified in believing P on the basis of a speaker's testimony, and is further led by this speaker to mistrust the testimony of other speakers whose testimonies contradict P, it is not only the hearer's future epistemic environment that comes under threat, but also her capacity to empathise with the testimony of those speakers who present evidence contrary to P. This manoeuvre is called *evidential preemption* (Begby, 2021).

tal interactions lack the "affective vividness" that one encounters in the offline world. Here, the design of social media is implicated: the amount of information conveyed by a potential target of your empathetic response on a social media platform can (in part) determine the quality of your empathetic response (Ventham, 2024). For example, if you come across a reel of a user who appears to be in an affective state, such as grief, your affective empathetic responses might be activated by this person's grief. However, you may need more information to correctly interpret that it is truly grief, and not something else, that the other person is experiencing (Mossner & Walter, 2024). Moreover, to *empathise well* and establish a moral connectedness to them, you may need to sit with your activated empathetic response to develop it further, rather than immediately scrolling onto the next video suggested by a personalised algorithm.

The perceptual quality of the information transmitted is indexed to the design features and affordances of a given platform—a user on X (formerly known as Twitter) can relay information about himself via tweets or posts that are limited to a certain number of characters, whereas on TikTok, he must make a reel to convey his subjective experience. In the case of a tweet, his thoughts and emotions are typically expressed through written words, hashtags, and emojis, which means that you do not have access to his voice, embodied expression, and other perceptual markers that can assist you in understanding his affective state. However, a video format has its own problems—while he is now visually accessible, he cannot be smelled, asked questions, or tapped reassuringly on the shoulder, and you only have access to him through the different frames of the video uploaded that may elide important background information (e.g., what is just outside the camera shot).

Compare this with face-to-face experiences, where you have embodied access to another person, which enriches your empathetic responses. If someone expresses grief in person to you, you have access not only to her speech but also to different perceptual markers (e.g., how she is holding her body, the tenor of her voice, her mood, etc.) that can assist you in the affective empathy process. You can also respond to her bodily movements with your own: if she moves forward to express dismay, you can receive her dismay and react to it in real time. You can show her that you care by expressing concern in your facial expressions and bodily movements, and by giving her your undivided attention. Face-to-face experiences are typically much denser in information and qualitatively richer, which can enhance both the cognitive capacities discussed above and our affective empathetic responses and expression (Torbjørnsen & Hipolito, 2024). Clearly, then, digital affective empathy is distinct from face-to-face affective empathy and comes with its own specific challenges and issues.

4 Moral Assessment of Digital Empathy

We now draw on our descriptive account of digital empathy to consider the moral issues it raises. We break these moral issues down into: (1) *one-off* and (2) *repeat* failures of digital empathy, with a particular focus on toxicity and harassment as this drives much of the polarisation discourse (Anderson, 2022); (3) the impact of algorithms *automating* our exposure to candidates for empathy by *obscuring* and

framing empathy targets, making it *harder* to empathise well, and *rushing* empathetic engagement; and (4) the cumulative negative effects this could have on our *moral skills*. Of course, this leaves open the possibility that there are failures of digital empathy that are not morally problematic,[9] and that social media might improve or increase our empathy skills through exposure to greater diversity. However, given that the existing empirical evidence outlined above does not suggest that the impacts of social media on empathy are overwhelmingly positive, and that the distinctive philosophical questions arose primarily from how platform design and recommendation system generate novel *challenges* for empathising well, we focus here on its moral harms, without denying the possibility of moral benefits.

4.1 Morally Problematic One-off Failures of Digital Empathy

On our account, a *morally problematic one-off failure of empathy* refers to a specific one-off instance where the process of digital empathising with another *fails*[10], and this *results in harms*, including moral harms, such as cases where another's subjectivity is unreasonably regarded as unintelligible, leading to them being dismissed, ridiculed, or humiliated. These digitally mediated harms are commonplace and often suffered by those targeted by online hate speech, fighting words, propaganda, and derogatory speech (see Brison, 1998). Even interpersonal conflicts and breakdowns can result in moral harm of this form, especially when one feels *unseen and unheard* by another (Bailey, 2022). Empathising well can ensure that these harms either do not occur, can be reversed, or minimised, as demonstrated by our earlier wheelchair example.

Furthermore, empathising well is a skill that requires practice and self-regulation of your empathetic responses. As we discussed in the previous section, the digital environment presents its own distinct challenges and risks that impact the exercise of this skill. Failures of digital empathy occur for many reasons, such as a lack of adequate information, misidentification of another person's responses, biases and stereotypes that affect how we perceive their credibility, or a lack of necessary skills to empathise effectively (e.g., young children). Empathy failures can track either affective or cognitive components of empathy (or both). We delineate and analyse this by the following example of a one-off failure of digital empathy. While this example is imaginary, it is inspired by real cases of online harassment, as we note below. Furthermore, the aim of this example is to *illustrate* the mechanics of empathy failures that our conceptual framework identifies (such as the misidentification of affect and the framing role of recommendation algorithms), rather than serve as evidence for that framework.

[9] We will only focus on failures of empathy that may be morally problematic. We will thus not analyse *all* instances of failures of digital empathy (e.g., failing to empathise with a football fan's celebration of her team's win on YouTube), since we are only interested in capturing instances that raise moral concerns. As such, references below to "failures of empathy" should be taken as shorthand for morally problematic instances of this.

[10] We understand failures of empathy to be varied, since empathy can go wrong in multiple ways. Failure to empathise is one such way. One could also empathise too much, too little, or inaptly, as we explore through different examples below.

Consider the example of Jude, who is ideologically committed to conservative values and firmly believes that progressives are leading an assault on traditional values. His favourite influencer, Irvin, shares a video by a progressive influencer named Jane. In this video, Jane discusses instances of experiencing sexual harassment in workplace settings. Irvin's post, however, states that he distrusts Jane's testimony, and he cracks a sexist joke about women misreading innocuous flirting as harassment. Jude agrees with Irvin and finds Jane's testimony annoying and bemusing. He imagines himself making fun of her and posts a sexist comment, before scrolling further down his feed. Assume that Jude often does not come across testimonies of sexual harassment in offline settings and moves through a conservative social scene that excludes feminist modes of thinking (i.e., he is stuck in a filter bubble, both online and offline). In his offline environment, he is unlikely to encounter such testimonials. However, in his online environment, he incidentally does so in this case. While technically he is exposed to more diverse information sources than usual, his response lacks empathy and care.[11] Why has he been unable to exercise empathy well in this case?

Many things have gone wrong. First, he does not self-regulate his empathetic response. As someone who is not used to feminist sources of information, for him to empathise well, he would have to up-regulate his response to Jane's testimony: i.e., he would have to pay careful attention to her testimony, make the effort to see the world from her point of view, and assess the morally salient features that arise from it. He would also need to make a concerted effort to interpret Jane's affective response: instead of attributing the misogynistic stereotype of 'attention seeking' to her, he should interpret her to be experiencing a more apt emotion (e.g. anger). By misinterpreting Jane's affective response, his own affective response to her of bemusement and annoyance further hinders him from imagining himself in her shoes. Consequently, he leaves a distasteful comment on Jane's post. He has thereby failed to empathise with Jane. Had he empathised with her, he could have gained more understanding of women's experiences in workplace settings. He could have still disagreed with some parts of Jane's claims (e.g., she argues that all workplace flirting is inappropriate), without failing to apprehend the reasons that animate Jane's perspective. In this instance, his failure of empathy encompasses both cognitive and affective empathy failures.

Part of the reason Jude has incurred a one-off failure of digital empathy in this example is due to his interpretation of Jane's testimony through a sexist normative landscape where women are considered inferior, infantile, and subordinate (Manne, 2017). He has employed stereotypes and social scripts that deny Jane her dignity and agency, undermining her testimony. Undoubtedly, these structural conditions of misogyny influence his response. However, additionally, another relevant contributory factor in this case is the role of social media in facilitating this failure. Jude encounters Jane's testimony through a post shared by Irvin, where Irvin has trashed

[11] In the background, we are drawing upon Nguyen's distinction between echo chambers and epistemic bubbles to inform our argument. An echo chamber, unlike a filter bubble, is much more difficult to escape. Exposure to new and diverse information, on its own, is not sufficient to challenge one's echo chamber membership. One has to reboot one's entire epistemic system to escape an echo chamber. See Nguyen (2020) on this distinction.

Jane's testimony using misogynistic tropes, and this has been algorithmically recommended to Jude by the platform.[12]

Exposing users to diverse content is only beneficial insofar as the platform's moderation system can effectively handle that diversity. When content moderation systems fail, issues of social media's responsibility towards its users arise. For example, a quick glance at the literature on social media design and recommender systems suggests that social media platforms are conducive to hostile and non-deliberative environments (see, e.g., Vaidhyanathan, 2018; Forestal, 2020, 2022). A consequence of causally bringing together diverse users—both by recommending users and their generated content to other users, and by providing them with design features to interact in some ways but not others (Lazar, 2023)—is that users will sometimes struggle to express empathy with one another in these online spaces, especially when they should attempt to, as is the case with Jude. At the very least, users should not be able to act with unjustified hostility to one another. Social media design and recommendation are therefore causally and, in part, morally responsible for Jude's failure of empathy by recommending to him a post marred by Irvin's misogynistic presentation of Jane's experiences. It is important to note here that the platform's responsibility is not merely that of providing a space where empathy failures might occur. Rather, the platform bears a distinctive form of responsibility because its recommendation systems actively directed Jude's attention toward Irvin's misogynistic reframing of Jane's testimony, thereby structuring the conditions under which his empathy failed. Of course, we must stress that this is not to deny that Jude is also morally responsible for his own behaviour and that for his response to change, patriarchal forces will also have to be challenged. But it is important to stress that social media moderation has played a role here in facilitating and encouraging Jude's moral failure. After all, social media sets the terms, conditions, and environment in which users connect.

Above, we have focused on a case of online harassment to express an instance of a morally problematic one-off failure of empathy. Unfortunately, these cases are not uncommon (Suzor, 2019). Moreover, one agent's non-empathetic response can often be sufficient to harm another agent, especially if the harm tracks existing structures of subordination and domination. By providing users with a digital environment where they can act in a hostile and aggressive manner to one another, and sometimes even rewarding that through recommendations, social media contributes to the harm that users cause to each other.

4.2 Morally Problematic Aggregate Failures of Empathy

Morally problematic aggregate failures of empathy refer to an aggregation of one-off morally problematic failures of digital empathy that often result from their amplification by recommendation systems. This warrants separate consideration, as aggregate failures that harm a marginalised person—such as Jane in the example above—often also create and contribute to a distinct ecological effect that goes

[12] This example is also an instance of evidential preemption. Followers of Irvin, in our example, are systematically told not to trust testimonies of feminist voices (such as Jane's), and those feminist voices are portrayed as untrustworthy.

beyond the aggregate harm experienced by that one individual. The distinctive ecological effect is that they create and promote digital environments that tend to be hostile and subordinating to often-vulnerable individuals and groups, especially minorities (see, e.g., Phillips, 2018). To illustrate this point, we expand our example from the previous subsection. Suppose Jude's distasteful comment on Jane's post starts receiving traction from others. After all, Jude has signalled to others his values, and by extension, he has invited them to either agree or disagree with his comment. Some might disagree with Jude and defend Jane's testimony.[13] But many more might agree with Jude and pile on by commenting in similarly distasteful ways. Since the distasteful comment is prompting further engagement, the recommendation algorithm might *amplify* this comment even further by making it visible to more people. As a result, Jane ends up experiencing virulent sexism from many users in the comments section of her post.[14] The recommendation algorithms are implicated: based on users' engagement with Jude's comment, the recommenders have directed more users' attention to Jude's explicit failure of empathy. By scaling up the visibility of Jude's comment, the recommendation systems have also prompted other users to join in, either by defending Jane or by expressing harmful speech against Jane. If this type of harmful speech had been demoted or Jude's comment censored in the first place, the likelihood of harm befalling Jane could have been reduced. Recommenders have *failed* to protect Jane and *added* to the harm other users have caused her by encouraging aggregate failures of empathy, while also contributing to an ecological effect of misogyny in online spaces.

The case as described, is unfortunately, common. When social media recommenders amplify or inadequately regulate (i.e. demote or censor) failures of empathy on their platforms, aggregate harms can emerge. Social media can become *lawless* (Suzor, 2019). Platforms are infamous for toxicity towards women and espousing anti-feminist values (see Richardson-Self, 2021, for examples). Women experience failures of empathy and their effects (i.e. harmful speech), in part due to bad content moderation practices. This results in aggregate harms against women, but similar trends have also emerged against ethnic and racial minorities, and LGBTQIA+ identifying people. This can create a digital environment that is hostile to members of these groups.

The role of bad content moderation practices in normalising harmful social norms must be emphasised. When failures of empathy result in harmful speech that is amplified across a platform, it indicates to users that community guidelines can be breached without consequence (if they are indeed breached). Community guidelines are the values that a platform promotes for its users to follow when posting and interacting with one another on the platform (Gillespie, 2018). On most large platforms, a core value enshrined in these guidelines emphasises that users treat one another with respect and refrain from engaging in harmful speech

[13] While people may come to the aid of others in digital spaces, this aid often happens after the abuse and trolling has already started and done much of its damage.

[14] Online abuse starts and amplifies in similar ways (see Barnes, 2023 for a philosophical analysis of online trolling and abuse as collective subordinating speech acts). For a recent example, see Dixon's (2024) discussion of her experience as a philosophy academic being subject to "r*pe threats" on X from a "light-hearted Hume Tweet".

(see, e.g., Meta's Community Standards).[15] When this value is violated and the violation is amplified, it signals to users, among other things, that the platform struggles to moderate content and the social norms enforced on it permit hostile behaviour. When such norms are *actualised* and *normalised* in online environments through aggregate harms, this can signal to minority users, such as women, that failures of empathy resulting in harmful speech are tolerated on this platform, which makes it potentially unsafe for them to express themselves, thereby excluding and silencing them and their experiences from that platform.

4.3 The Automation of Digital Empathy

Recall that recommenders are "the algorithms driving social media" (Narayanan, 2023, p. 5), responsible for displaying, curating, and recommending content to users. In our digital environments, almost everything we encounter has been predetermined by platforms (Gowder, 2023). If a platform determines that it is suboptimal for a user to discover news about, for example, the Black Lives Matter movement on its platform, then it can make that the case, irrespective of the user's personal tastes, values, desires, and beliefs. Recommendation systems thereby select our *candidates for engagement* on a platform (i.e., what users and user content we see), and update those candidates in real-time by observing our behaviour on the platform (see Sect. 2.2). This impacts digital empathy in four important ways: by *obscuring* targets of empathy, by *framing* targets of empathy, by making it *harder* to empathise well, and by *rushing* empathetic engagement.

First, recommendation systems can *obscure* genuine targets of empathy by not recommending those targets.[16] For example, imagine John is politically neutral about refugees. On YouTube, John is mostly recommended videos about his favourite movies and the sports team that he follows, but he never encounters news videos about refugees. Consequently, a category of empathy targets (i.e. people John could be exposed to that may provoke empathetic responses in him) is never selected for engagement by YouTube's recommendation system. Second, recommendation systems can sometimes direct users to targets who are prominently featured in online spaces, but these targets are often portrayed in a negative light. This is a case of framing (Benkler et al., 2018). For example, on Instagram, John is directed to users who post derisive stories on their feeds about refugees. Through those stories, John learns about different refugee crises and encounters users' testimonies about refugees. However, since those testimonies are framed for John by other users' negative stereotypes about refugees as part of a divisive online environment, John must significantly up-regulate his empathetic response to move past those negative stereotypes about refugees if he is to empathise well with them. In both instances, the targets of empathy are either obscured by not being shown (as

[15] Of course, these guidelines change from platform to platform (see Gillespie, 2018 for empirical analysis of different platforms' community guidelines).

[16] This obscuring of genuine targets of empathy is not unique to social media platforms (Alfano & Skorburg, 2018). Unsurprisingly, many of our concerns can also be extended to traditional media outlets, although exploring this extension further is clearly beyond the scope of this paper.

in the YouTube example) or are framed by being negatively portrayed within the platform's ecosystem (as in the Instagram example).

Third, platforms can make empathising well *harder* by amplifying bad targets of empathy by recommending them across social networks. Let's return to our example from the previous subsection. Suppose that on Instagram, toxic men's rights activists are recommended to Jude. By recommending and amplifying toxic men's rights activists on Jude's feed, Instagram's recommendation systems direct his attention toward bad empathy targets (i.e., targets that are morally deficient or vicious). He consumes their content and starts empathising with their concerns, which leads him to develop more extremist misogynistic views towards women and feminism as a political movement. For Jude to empathise well with women, he will now have to both down-regulate his empathetic response with men's rights activists and up-regulate his empathetic response towards feminists. By getting affectively committed to empathising with men's rights activists through a recommendation system, it makes it harder for Jude to regulate his empathetic responses in morally appropriate ways.

Fourth, recommendation systems can *rush* a user's empathetic response. Recall that platforms commercialise attentional resources and are geared towards distraction (there is always something new to look at!) to promote more extended engagement. Users' attention is directed from one piece of content to another, often in rapid succession. For some platforms (e.g., TikTok), content of a short duration achieves most engagement, causing short content to be upranked in recommendations (Guinaudeau et al., 2022). Moreover, different types of affective content are often placed next to one another or follow quickly one after another, which can provoke various emotions in users. Recall that well-allocated empathy is focused and can be time-consuming, especially if self-regulation is involved. To empathise well, we must sit with our own feelings and critically regulate our empathetic responses to others. Imagining ourselves into another's perspective, accurately grasping its intelligibility, and sitting with the emotions it evokes takes both time and effort. But many platforms are designed to rush you along to the next piece of content to maintain platform engagement in a way that hinders empathising well. This raises broader issues about moral deskilling to which we now turn.

4.4 Moral Deskilling and Digital Empathy

Moral deskilling refers to the redundancy of a moral skill due to advancements in technology (Vallor, 2015). While virtue theorists have posited moral skill acquisition as necessary (though not sufficient) for the development of a virtuous character (see, e.g., Jacobson, 2005), we do not need a commitment to Aristotelian virtue ethics to see why moral skills are morally important. Moral skills can, among other things, help agents identify and respond to the morally salient features of a situation and explain their reasons for acting. Practising moral skills in relevant contexts allows agents to develop and cultivate those skills. As argued above, empathy is also a skill, and agents must practice this skill to become good at empathising. The moral deskilling of empathy refers to the diminishing of the skill of empathy across human society due to the advancement of social media (and other) technologies. As empathy becomes deskilled, humans will struggle

to practice it effectively, which could lead to a decline in moral connectedness. This is one part of a broader discussion around the links between automation, technological outsourcing, and deskilling (Vallor, 2016).

There are two key concerns here. First, the primary reason for worrying about the deskilling of empathy in general is due to the difficulties associated with practising digital empathy well on social media. To reiterate the claims that we have argued for thus far, digital empathy faces challenges on social media because platform design permits individual failures of empathy that lead to individual and aggregate instances of harm, and platform recommendation systems can obscure and frame genuine targets of empathy, amplify both failures of empathy and inappropriate targets of empathy (i.e. white supremacists), and rush empathetic responses in users. Second, our empathy skills may erode over time as we spend more and more time in digital environments, where it is difficult to practice digital empathy effectively for the reasons summarised below. As we continue to spend more time on social media (see Gottfried, 2024; Global Social Media Statistics, 2024, for the latest statistics on upward trends in social media use), the embodied and perceptual skills that we use in informationally rich and diverse face-to-face interactions to empathise well with others in offline contexts might start to atrophy. Social media platforms might also help to normalise failures of empathy, which could lead to a more polarising and divisive political culture of deliberation where we are less likely to self-regulate our empathetic responses. Such a trajectory is worrisome because empathy is valuable and serves important moral and epistemic roles.

While the empirical evidence concerning the negative impact of social media on empathy is mixed (as noted above), the concern of moral deskilling of empathy is still worth taking seriously. This concern is best framed as a present concern about a problem that may become more severe in the future, especially if the status of empathy in online spaces continues to worsen. Offline acquisition and development of the skill of empathy may, for now, mitigate some of the ill effects and challenges associated with digital empathy. However, as we continue to inhabit digital environments more frequently, the empathy skills we learn offline may not be sufficient to counter the empathy deskilling effects we encounter online. Furthermore, for minors and young adults, it may already be the case that due to an exposure to online environments from an early age, they are unable to develop the right kind of perceptual and embodied skills of picking up on the rich informational diversity we are exposed to in face-to-face interactions needed to empathise well (Mossner & Walter, 2024; Haidt, 2024). Instead of experiencing deskilling of an existing skill, in the case of children and adolescents, digital environments may interfere with the acquisition and development of the skill of empathy in the first place.

5 The Moderation of Empathy

Given the potential negative impacts on digital (and, by extension, overall) empathy, is there an obligation for platform owners, platform regulators, or platform users to address these impacts?[17] Many interventions have been suggested in the literature (e.g., Mossner & Walter 2024, 7–10; Haidt, 2024), ranging from offline behavioural interventions (e.g., meditation, reading, etc.), to automatic nudges on social media aimed at improving the capacity for digital empathy amongst internet users, to banning social media for teens and adolescents under 16. In this section, instead of summarising these interventions, our aim is to address whether social media platforms have a moral obligation to moderate digital empathy. By "moderating" digital empathy, we mean implementing mechanisms on a platform with the intentional goal of changing people's experiences of digital empathy. Of course, social media platforms already implicitly impact the digital empathy experiences of their users as a side effect of their existing architecture and recommendation algorithms, as we have seen above. But is explicitly moderating the digital empathy experiences of users, in the same way that certain forms of speech and other content, such as hate speech and misinformation, are explicitly moderated on platforms, a morally worthy or required goal?

The existing literature offers several frameworks for justifying content moderation. Howard (2024) proposes what we can term the 'Complicity Argument', arguing that social media companies have a moral duty to avoid complicity in the harms caused by users' speech. This argument justifies the use of content moderation systems capable of censorship, demotion, and amplification. Similarly, Kashyap and Formosa (2023) argue that the private property rights of platform owners justify their power of content moderation; however, the exercise of these powers is limited when it results in significant harm to others or the perpetration of injustices. These approaches primarily focus on preventing harms and injustices, which aligns with the traditional justifications for content moderation (Keller, 2021).

Applying this harm-prevention framework to digital empathy, we can make a prima facie case for the moral obligation to moderate content to limit what we term "empathy harms." We define empathy harms as negative consequences resulting from failures of digital empathy that undermine moral connectedness between users. These harms can be categorized into four types, as outlined earlier: (1) One-off failures: for example, a user posting a dismissive comment on a heartfelt post about personal loss, failing to recognize the poster's emotional state and causing them harm as a result; (2) Aggregate harms: such as the cumulative effect of multiple users engaging in empathy failures, creating a hostile environment for vulnerable groups; (3) Algorithmic harms, including obscuring and framing targets of empathy, making it harder to empathise well, and rushing empathetic engagement; (4) Empathy deskilling: The gradual erosion of empathy skills due to prolonged exposure to digital environments that do not adequately support

[17] Of course, empathy harms that track structures of domination and subordination cannot simply be addressed by fixing up platform architecture and content moderation operations. To address these harms, we will also need to directly challenge harmful ideologies, such as sexism and racism.

empathetic engagement. This argument can ground a moral duty on social media companies to moderate content to minimise (if not eliminate) empathy harms.

At this stage, it is useful to situate the harm-prevention justification for moderating empathy harms within the broader platform governance literature. Three prominent models are worth briefly highlighting. First, consider speech-based harm models (see, e.g., Ganapini, 2026), which prioritise the regulation of harmful speech, and which may be justified by appeals to the Complicity argument (Howard, 2024). Second, democratic models emphasise that platforms should govern speech in accordance with values such as explainability, accountability, and transparency (see, e.g., Suzor, 2019; Kashyap & Formosa, 2023). Third, market-oriented approaches emphasise user choice and competition as mechanisms for constraining platform power (Keller, 2021).

While each of these models captures an important dimension of platform governance, the moderation of empathy presents unique challenges that are not fully addressed by any of them. Speech-based harm models target clear-cut cases of harmful content, such as hate speech. However, many empathy failures do not result from or involve speech that would typically be flagged by such models, such as a dismissive comment or a rushed scroll past someone's distress. Democratic models emphasise explainability and transparent moderation guidelines, but empathy is deeply contextual. Whether a given interaction constitutes a failure of empathy depends on interpersonal dynamics, cultural context, and subjective emotional states that resist codification into clear and transparent rules. Market-oriented models rely on user choice and competition to constrain platform power, but empathy deskilling is a gradual harm that individual users may neither notice nor be able to escape through switching platforms. Moreover, as Mossner and Walter (2024) highlight, the difficulties of fostering empathy in digital spaces stem from structural features, such as the lack of perceptual depth and diversity compared to face-to-face interactions, that cannot be removed through better moderation.

The literature on algorithmic governance and content regulation provides further insights into the challenges of moderating for empathy. Narayanan (2023) highlights the potential for unintended consequences when attempting to shape user behaviour through platform design and moderation. In the context of empathy moderation, overzealous efforts to prevent empathy harms could lead to chilling effects on free speech, potentially undermining the very goal of fostering empathy by limiting diverse perspectives, genuine but challenging social interactions, and authentic emotional expressions. Further, efforts to moderate empathy are complicated by cultural differences in empathetic expression and interpretation. Concerns such as this lead Gorwa et al. (2020) to argue that optimised moderation systems might exacerbate rather than relieve existing content moderation problems by increasing the opacity of moderation systems and obscuring foundational political issues leading to "de-politicisation".

The distinction between preventing harm and promoting moral improvement further complicates the issue of empathy moderation. This distinction is grounded in philosophical debates about the limits of paternalism and the role of institutions in moral development. Mill's harm principle suggests that preventing harm to others is the only legitimate reason for interventions that restrict liberty, whereas

efforts to morally improve individuals are more contentious (Arneson, 2005). In the context of digital empathy, while harm prevention aligns with established justifications for moderation, actively seeking to improve users' empathy skills ventures into ethically controversial territory. It is one thing for a user of a platform to seek to intentionally cultivate their digital empathy skills while using digital tools, and another thing for an opaque recommendation system to seek to nudge them automatically without their knowledge towards a similar goal. This echoes broader concerns in the philosophy of technology literature about the ethical implications of persuasive design and nudging in digital environments (Benner et al., 2022).[18] The paternalism concern is particularly acute when considering empathy moderation, as it involves shaping users' emotional responses and interpersonal skills, which are areas traditionally considered private and personal, although they clearly also have a political impact on the quality of civic interpersonal connections.

Given these complexities, it remains unclear whether and to what extent social media platforms should explicitly moderate content to foster digital empathy. While there seems to be a reasonable case for moderating to prevent clear empathy harms based on existing justifications for content moderation, the potential drawbacks and ethical concerns raised above suggest the need for caution. Considering how social media companies could optimize recommenders to improve users' empathy targets, experiences of digital empathy, and users' digital empathy skills will require addressing difficult questions regarding the personalisation of the public sphere (Forestal, 2020), the moral and political values we ought to prioritize in ranking online content (Lazar, 2023), paternalism worries, increased algorithmic opacity concerns (Gorwa et al., 2020), and the free speech costs of demoting content that is deemed harmful. Future research should focus on developing nuanced frameworks for understanding the ethical implications and trade-offs of explicitly moderating digital empathy, the moral costs of inaction, and the potential of moderation technologies that could foster more empathetic online environments while respecting user autonomy and free expression. Furthermore, addressing the challenges of digital empathy may require a multifaceted approach that extends beyond content moderation alone, such as educational initiatives to enhance digital literacy and empathy skills (Ventham, 2024), as well as broader efforts to reduce the toxicity of offline political and civic discourse.

6 Conclusion

In this paper, we presented a framework of digital empathy, beginning with a descriptive account that encompassed both affective and cognitive components. We then explored the impacts of platforms on digital empathy and offer four main categories of moral assessment: (1) morally problematic one-off failures of digital empathy; (2) aggregate harms of failures of digital empathy; (3) the impacts of recommendation systems on digital empathy by amplifying, obscuring, framing, and rushing targets of empathy; and (4) the deskilling of empathy due to the widespread

[18] Not everyone thinks that nudging is a problem (see, e.g., Levy, 2019).

use of social media. We also explored harm-based justifications for moderating content to explicitly govern empathetic relationships between users of a platform. By taking our argument in this direction, we aim to connect discussions of digital empathy with the wider literature on the ethics of content moderation and encourage others to think further about the moral, social, and political implications of social media's role in connecting people and facilitating their interactions, including its impacts on empathy.

Acknowledgements The authors would like to thank Neil Levy, Mark Alfano, Paul-Mikhail Catapang Podosky, and Tom Montefiore for comments and feedback.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. Bhanuraj Kashyap's research is funded by an International Macquarie University Research Excellence Scholarship (Scholarship Allocation Number: 20224272).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Alfano, M., Joshua, A., & Skorburg (2018). Extended knowledge, the recognition heuristic, and epistemic injustice. In *Extended Knowledge*, edited by D. Pritchard, J. Kallestrup, O. Palermos, & A. Carter, 239–256. Oxford University Press.

Anderson, E. S. (2022). Can we talk? Communicating moral concern in an era of polarized politics. *Journal of Practical Ethics*, *10*(1). https://doi.org/10.3998/jpe.1180

Andrejevic, M. (2014). *Infoglut*. Routledge.

Arguedas, R., Amy, C. T., Robertson, R., Fletcher, & Rasmus Kleis, N. (2022). Echo chambers, filter bubbles, and polarisation. *Reuters Institute*. https://doi.org/10.60625/RISJ-ETXJ-7K60

Arneson, R. J. (2005). Joel Feinberg and the justification of hard paternalism. *Legal Theory*, *11*(3), 259–284. https://doi.org/10.1017/S1352325205050147

Avenanti, A., & Sirigu, A., and Salvatore M. Aglioti (2010). Racial bias reduces empathic sensorimotor resonance with other-race pain. *Current Biology*, *20*(11), 1018–1022. https://doi.org/10.1016/j.cub.2010.03.071

Bailey, O. (2022). Empathy and the value of humane understanding. *Philosophy and Phenomenological Research*, *104*(1), 50–65. https://doi.org/10.1111/phpr.12744

Barnes, M. (2023). Whom do you speak for? And How? *Journal of Ethics and Social Philosophy*, *25*(2). https://doi.org/10.26556/jesp.v25i2.1561

Barrett, D. (2022). Political polarization and social media. *Philosophical Topics*, *50*(2), 85–104. https://www.jstor.org/stable/48752501

Begby, E. (2021). Evidential preemption. *Philosophy and Phenomenological Research*, *102*(3), 515–530. https://doi.org/10.1111/phpr.12654

Benkler, Y., Farris, R., & Roberts, H. (2018). *Network propaganda*. Oxford Publishing.

Benn, C., and Seth Lazar (2022). What's wrong with automated influence. *Canadian Journal of Philosophy*, *52*(1), 125–148. https://doi.org/10.1017/can.2021.23

Benner, D., Schöbel, S., Janson, A., & Leimeister, J. M. (2022). How to achieve ethical persuasive design. *AIS Transactions on Human-Computer Interaction*, *14*(4), 548–577. https://doi.org/10.17705/1thci.00179

Brison, S. J. (1998). Speech, harm, and the mind-body problem in first amendment jurisprudence. *Legal Theory*, *4*(1), 39–61. https://doi.org/10.1017/S1352325200000914

Cohen, J., Fung, A. In digital technology and democratic theory, edited by Bernholz, Lucy, Hélène Landemore, and, & Reich, R. (2021). 23–61. University of Chicago Press. https://doi.org/10.7208/chicago/9780226748603.001.0001

Cuff, B. M. P., Sarah, J., Brown, L., Taylor, & Howat, D. J. (2016). Empathy: A review of the concept. *Empathy Emotion Review* 8 (2): 144–153. https://doi.org/10.1177/1754073914558466.

Darwall, S. (1998). Empathy, sympathy, care. *Philosophical Studies*, *89*, 261–282. https://doi.org/10.1023/A:1004289113917

Denham, A. E. (2017). Empathy and moral motivation. In The Routledge Handbook of Philosophy of Empathy, edited by Heidi Maibom, 227–241. New York: Routledge, 2017. https://doi.org/10.4324/9781315282015-21

Dixon, D. (2024). Women Philosophers in the Twitter 'Manosphere' (or, That Light-Hearted Hume Tweet That Ended in R*pe Threats). The Philosophers' Cocoon. Accessed July 11, 2024. https://philosopherscocoon.typepad.com/blog/2023/05/women-philosophers-in-the-twitter-manosphere-or-that-light-hearted-hume-tweet-that-ended-in-rpe-thre.html

Forestal, J. (2020). Constructing digital democracies. *Political Studies*, *69*(1), 26–44.

Forestal, J. (2022). *Designing for democracy*. Oxford University Press.

Formosa, P. (2017). *Kantian ethics, dignity and perfection*. Cambridge University Press.

Fricker, M. (2007). *Epistemic injustice*. Oxford Academic.

Ganapini, B., & Marianna (2026). Regulating misinformation online: permissible content moderation. *Inquiry: A Journal Of Medical Care Organization, Provision And Financing*. https://doi.org/10.1080/0020174X.2026.2624389

Gillespie, T. (2018). *Custodians of the internet*. Yale University Press.

Gillespie, T. (2022). Do not recommend? Reduction as a form of content moderation. *Social Media + Society*, *8*(3), 205630512211175. https://doi.org/10.1177/20563051221117552

Global Social Media Statistics. DataReportal: Global digital insights. Accessed July 15 (2024). https://datareportal.com/social-media-users

Goldman, A. (2006). *Simulating minds*. Oxford University Press.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation. *Big Data & Society*, *7*(1), 2053951719897945. https://doi.org/10.1177/2053951719897945

Gottfried, J. (2024). Americans' Social Media Use. Pew Research Center (blog). January 31, 2024. https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/

Gowder, P. (2023). *The Networked Leviathan*. Cambridge University Press.

Guan, S. S., Angie, S., Hain, J., Cabrera, & Rodarte, A. (2019). Social media use and empathy. *Social Networking*, *08*(04), 147–157. https://doi.org/10.4236/sn.2019.84010

Guinaudeau, B., & Munger, K., and Fabio Votta (2022). Fifteen seconds of fame. *Computational Communication Research*, *4*(2), 463–485. https://doi.org/10.5117/CCR2022.2.004.GUIN

Haidt, J. (2022). Why the past 10 years of American life have been uniquely stupid. The Atlantic. Accessed July 4, 2024. https://www.theatlantic.com/magazine/archive/2022/05/social-media-democracy-trust-babel/629369/

Haidt, J. (2024). *The anxious generation*. Random House.

Hall, J. (2020). and Mark Leary. The U.S. Has an Empathy Deficit. Scientific American. Accessed July 4, 2024. https://www.scientificamerican.com/article/the-us-has-an-empathy-deficit/

Hindman, M. (2018). *The internet trap*. Princeton University Press.

Hirji, S. (2022). Outrage and the bounds of empathy. *Philosophers' Imprint*, *22*(0). https://doi.org/10.3998/phimp.1296

Howard, J. (2024). The ethics of social media: Why content moderation is a moral duty. *Journal of Practical Ethics*, *11*, 33–52.

Hume, D. (1975). *Enquiries concerning human understanding and concerning the principles of morals*. Clarendon.

Husserl, E. (1970). *The cartesian Meditations*. Martinus Nijhoff.

Jacobson, D. (2005). Seeing by feeling. *Ethical Theory and Moral Practice*, *8*(4), 387–409. https://doi.org/10.1007/s10677-005-8837-1

Kant, I. (1996). The Metaphysics of Morals. In Practical Philosophy, edited by M.J. Gregor, 353–604. Cambridge University Press.
Kashyap, B., and Paul Formosa (2023). The authority to moderate: Social media moderation and its limits. *Philosophy & Technology*, *36*(4), 78. https://doi.org/10.1007/s13347-023-00685-w
Kauppinen, A. (2014). Empathy, emotion regulation, and moral judgment. In Empathy and Morality, edited by Heidi L. Maibom. Oxford University Press.
Kauppinen, A. (2017). Empathy and moral judgment. In The Routledge Handbook of Philosophy of Empathy, edited by Heidi L. Maibom, 215–226. Routledge.
Keller, D. (2021). *Amplification and its discontents*. Knight First Amendment Institute.
Kennett, J. (2017). Empathy and psychopathology. In The Routledge Handbook of Philosophy of Empathy, edited by Heidi L. Maibom, 364–376. Routledge.
Konrath, S. (2013). The empathy paradox. In Handbook of Research on Technoself, edited by Rocci Luppicini. IGI Global.
Konrath, S., Martingano, A. J., & Davis, M., and Fritz Breithaupt (2023). Empathy trends in American youth between 1979 and 2018. *Social Psychological and Personality Science*. https://doi.org/10.1177/19485506231218360
Lazar, S. (2023). Communicative justice and the political philosophy of attention. Accessed 28 Jul 2023. https://hai.stanford.edu/events/tanner-lecture-ai-and-human-values-seth-lazar
Levy, N. (2019). Nudge, nudge, wink, wink: Nudging is giving reasons. *Ergo an Open Access Journal of Philosophy*, *6*. https://doi.org/10.3998/ergo.12405314.0006.010
Maibom, H. (2007). The presence of others. *Philosophical Studies*, *132*, 161–190. https://doi.org/10.1007/s11098-004-0018-x
Maibom, H. L. (2014). Introduction: (Almost) Everything you ever wanted to know about empathy. In Empathy and Morality, edited by Heidi L. Maibom, 1–40. Oxford University Press.
Maite Taboada. (2021). The banality of online toxicity. Policy Options. Accessed July 4, 2024. https://policyoptions.irpp.org/magazines/october-2021/the-banality-of-online-toxicity/
Manne, K. (2017). *Down girl*. Penguin Books.
Mossner, C., & Walter, S. (2024). Shaping social media minds: Scaffolding empathy in digitally mediated interactions? *Topoi*, *43*(3), 645–658. https://doi.org/10.1007/s11245-024-10034-x
Narayanan, A. (2023). Understanding social media recommendation algorithms. Knight First Amendment Institute. Accessed July 4, 2024.
Nguyen, C., & Thi (2020). Echo chambers and epistemic bubbles. *Episteme*, *17*(2), 141–161. https://doi.org/10.1017/epi.2018.32
Pariser, E. (2011). *The filter bubble*. Penguin Books.
Phillips, W. (2018). *The oxygen of amplification*. Data and Society.
Prinz, J. (2011a). Is empathy necessary for morality? In *Empathy: Philosophical and Psychological Approaches*, edited by A. Coplan & P. Goldie, 211–229. Oxford University Press.
Prinz, J. (2011b). Against empathy. *The Southern Journal of Philosophy*, *49*(s1), 214–233. https://doi.org/10.1111/j.2041-6962.2011.00069.x
Ranalli, C., & Malcom, F. (2025). What's so bad about echo chambers? *Inquiry*, *68*(10), 3984–4026. https://doi.org/10.1080/0020174X.2023.2174590
Richardson-Self, L. (2021). *Hate speech against women online*. Rowman & Littlefield.
Rubin, J. (2024). An empathy deficit explains many of our problems. The Washington Post. Accessed July 4. https://www.washingtonpost.com/opinions/2023/02/15/empathy-gap-fix/
Schramme, T. (2024). Empathy as a means to understand people. *Philosophical Explorations*, *27*(2), 157–170. https://doi.org/10.1080/13869795.2024.2344975
Slote, M. (2010). *Moral sentimentalism*. Oxford University Press.
Smith, A. (1976). *The theory of moral sentiments*. Oxford University Press.
Srnicek, N. (2017). *Platform capitalism*. Polity.
Suzor, N. (2019). Lawless. Cambridge University Press.
Torbjørnsen, R., & Rørstad (2024). and Inês Hipólito. Widening the Screen. AI & SOCIETY. https://doi.org/10.1007/s00146-023-01844-5
Vaidhyanathan, S. (2018). *Antisocial media*. Oxford University Press.
Vallor, S. (2012). Flourishing on Facebook. *Ethics and Information Technology*, *14*(3), 185–199. https://doi.org/10.1007/s10676-010-9262-2
Vallor, S. (2015). Moral deskilling and upskilling in a new machine age. *Philosophy & Technology*, *28*(1), 107–124. https://doi.org/10.1007/s13347-014-0156-9
Vallor, S. (2016). *Technology and the virtues*. Oxford University Press.

Ventham, E. (2025). Empathy, motivating reasons, and morally worthy action. *The Journal of Value Inquiry*, *59*(4), 827–839. https://doi.org/10.1007/s10790-023-09972-3

Ventham, E. (2024). Empathising in online spaces. *Philosophical Explorations*, *27*(2), 225–236. https://doi.org/10.1080/13869795.2024.2344980

Vossen, H. G. M., and Patti M. Valkenburg (2016). Do social media foster or curtail adolescents' empathy? A longitudinal study. *Computers in Human Behavior*, *63*, 118–124. https://doi.org/10.1016/j.chb.2016.05.040

Vuorre, M., & Przybylski, A. K. (2023). Global well-being and mental health in the internet age. *Clinical Psychological Science*, *12*(5), 917–935. https://doi.org/10.1177/21677026231207791

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.