Philosophical Studies
https://doi.org/10.1007/s11098-023-02083-6

Attention, moral skill, and algorithmic recommendation

Nick Schuster[1] · Seth Lazar[1]

Accepted: 17 November 2023

Abstract
Recommender systems are artificial intelligence technologies, deployed by online platforms, that model our individual preferences and direct our attention to content we're likely to engage with. As the digital world has become increasingly saturated with information, we've become ever more reliant on these tools to efficiently allocate our attention. And our reliance on algorithmic recommendation may, in turn, reshape us as moral agents. While recommender systems could in principle enhance our moral agency by enabling us to cut through the information saturation of the internet and focus on things that matter, as they're currently designed and implemented they're apt to interfere with our ability to attend appropriately to morally relevant factors. In order to analyze the distinctive moral problems algorithmic recommendation poses, we develop a framework for the ethics of attention and an account of judicious attention allocation as a moral skill. We then discuss empirical evidence suggesting that attentional moral skill can be thwarted and undermined in various ways by algorithmic recommendation and related affordances of online platforms, as well as economic and technical considerations that support this concern. Finally, we consider how emerging technologies might overcome the problems we identify.

Keywords Artificial intelligence · Attention · Machine learning · Moral skill · Recommender systems

1 Introduction

As Nobel laureate Herbert Simon first noted over 40 years ago, a wealth of information makes attention a scarce resource (Simon, 1971). Since then, the rise of the internet, big data, and artificial intelligence (AI) has not only drastically increased the amount of information available to us, it's also fundamentally transformed how

✉ Nick Schuster
nick.j.schuster@gmail.com

Seth Lazar
seth.lazar@anu.edu.au

[1] The Australian National University, Canberra, Australia

Published online: 22 January 2024

information is produced, distributed, and consumed. This technological revolution has therefore had a transformative effect on the attention economy.[1]

Among the most notable features of the digital attention economy is the way online platforms filter and rank information according to individual users' preferences, among other factors. This process centrally involves algorithmic recommender systems. Built with advanced machine learning (ML) techniques, these tools now play a critical role in directing our attention in digital environments to the products we buy, the entertainment we consume, the news we read, and even the jobs we work, the homes where we live, and the people with whom we interact. Indeed, the information saturation of the digital world has made recommender systems not just useful but necessary for effective agency.

These tools, and the sociotechnical systems in which they're embedded, therefore raise pressing moral concerns about how we now allocate our attention. On the one hand, they enable us to access an unprecedented amount of information with ease, potentially augmenting our agency in morally significant ways. For instance, they can direct us to far more information of personal, social, and political importance than we would be exposed to without them. On the other hand, our reliance on algorithmic recommendation could reshape our agency in morally problematic ways. Online platforms can direct us toward things we should not attend to just as easily as toward things we should. And even when they direct our attention to the right things, they may not do so in the right ways, to the right degrees, or for the right reasons.

In this essay, we argue that judicious attention allocation is a moral skill and that our growing reliance on algorithmic recommendation threatens our development and exercise of that skill. In Sect. 2, we situate our account of judicious attention allocation in a broader theory of the ethics of attention. In Sect. 3, we draw on empirical evidence and technical considerations to argue that recommender systems and the online platforms they serve can thwart and undermine attentional moral skill in various ways. Finally, in Sect. 4, we consider how generative AI systems could filter and rank content such that they scaffold and protect judicious attention allocation instead. A brief conclusion follows.

2 Judicious attention allocation as a moral skill

2.1 The ethics of attention

In the broadest terms, attention is "the selective directedness of our mental lives" (Mole, 2021, p. 1). While we can be generally aware of multiple things at once, we can only focus our attention on a relatively small subset at any given time. Attention

[1] Critics have raised a multitude of worries about AI technologies in the attention economy in recent years: AI might weaken market competition and democratic governance (Hindman, 2018), lead to social isolation (Turkle, 2011) and addiction (Bhargava & Velasquez, 2021), enable exploitation (Bueno, 2016), and erode individual liberties (Williams, 2018), to name a few. The concerns we'll raise are compatible with these worries, but don't depend on any of them.

can be fleeting, or it can be more sustained. Momentarily noticing a headline as you scroll through your newsfeed and getting engrossed in a story are both ways of paying attention. Attention can also be exogenous (automatically responsive to stimuli) as well as endogenous (deliberately directed by the agent). An alarm can grab your attention without you having to think about it or make an effort to focus on it; but if you know it's coming you can listen for it and once you hear it you can actively resist distractions.

Insofar as effective action requires us to focus on certain features of our circumstances while being at most peripherally aware of others, allocating our attention well is necessary for good practical agency (Allport, 1987; Neumann, 1987; Watzl, 2017; Wu, 2011). If you don't pay attention to the time, you'll miss your three o'clock meeting. And regular deficient attention to such things could jeopardize your career. Indeed, attention plays a central role in explaining various aspects of practical agency, including perception (Carrasco, 2011), emotion (Brady, 2013; De Sousa, 1990), decision-making (Orquin & Loose, 2013), intentionality (Wu, 2011, 2016), and self-control (Berkman et al., 2017; Bermúdez, 2017).

As such, attention allocation is subject not just to descriptive analysis but to normative assessment as well. There are, for instance, better and worse ways of allocating our attention in the service of both knowledge acquisition and practical ends. Thus, there are both epistemic and prudential norms of attention. Checking your calendar is a *good* way to verify the location of your meeting; and you *should* watch for traffic as you walk there, assuming you want to arrive in one piece. Moreover, if you've promised to be there, your attention allocation is also subject to moral norms. It's not just epistemically and prudentially good to pay attention to your schedule and the traffic, it's morally good as well.

Noting that attention allocation is normatively assessable in these ways, Sebastian Watzl (2022) has recently proposed a framework for the ethics of attention which "classifies potential norms of attention along three dimensions: whether they are manner or object-based, instrumental or noninstrumental, and whether its source is moral, prudential or epistemic" (p. 99). We agree that a multidimensional approach is necessary for working out the ethics of attention, but we propose a different framework. The primary task for the ethics of attention, as we understand it, is to morally assess attention allocation. We therefore take the primary dimensions to be those along which attention allocation can be morally better or worse. And so the framework we propose differs from Watzl's in two key respects.[2]

First, while we agree with Watzl that a full theory of the ethics of attention should explain how prudential, epistemic, and other non-moral norms interact with moral ones (Watzl, 2022, p. 90), we treat such questions as secondary. This is because attention allocation can be morally good or bad independently of whether it satisfies other kinds of norms. Spying on your ex, for instance, can be epistemically good in that it can justify your belief that she's now dating that "friend" of hers you always

[2] This isn't meant as a critique of Watzl, however, since he may not share our purposes. We develop our framework specifically to guide moral assessment of attention allocation and ultimately to ground our sociotechnical critique of algorithmic recommendation.

suspected. And reading leaked emails to engage in insider trading can be prudentially good in that it can promote your own financial interests. But these are both cases of morally bad attention allocation. Whether norms of attention are prudential, epistemic, or moral (or rational or aesthetic, for that matter) therefore isn't itself a dimension of moral assessment.

Second, whereas the object/manner distinction is a single dimension of Watzl's framework, we take the object of attention and the manner of paying attention to be two separate dimensions of moral assessment. You can attend to the right things yet not in the right ways. Moreover, we separate the manner of attention allocation into two further independent dimensions: degree and kind. Whether you pay attention to something to the right degree is a separate question from whether you attend to it through the right sort of psychological process, for instance. Finally, the reasons for which you attend as you do are morally assessable independently of all these considerations. Thus, we propose four distinct (primary) dimensions along which attention allocation can be morally better or worse: (1) which things one attends to; (2) how much one attends to different things; (3) the kind of attention one gives them; and (4) why one allocates one's attention this way. To illustrate how each of these dimensions matters independently, consider a well-worn case.

You see a small child struggling in shallow water. You should pay attention to this situation. But just attending isn't enough, you must also pay *enough* attention. If you're so focused on your own lap time that you respond slowly and distractedly, this is morally problematic. You attend to the right object yet not to the right degree. And even if you do pay enough attention, if you have to make a deliberate effort to stay focused, this too is morally suboptimal. All else being equal, it reflects better on you if the situation exogenously grabs and holds your attention. Finally, even if you pay the right kind of attention, in the right degree, to the right things here, you might not do so for the right reasons. If you attend to the drowning child only because you're babysitting her, for instance, and you fear legal consequences, this too would be less than ideal.[3]

Thus, what you attend to, how much, in what way, and for what reasons all come to bear independently on the moral assessment of your attention allocation. Moral philosophers have recently explored how attention functions for certain moral virtues (Bommarito, 2013; Chappell & Yetter-Chappell, 2016) and the role it plays in moral perception (Vance & Werner, 2022; Waggoner, 2021). Additionally, we submit that judiciously allocating attention can fulfill duties of attention. Some duties of attention are derivative from duties of action. To save the child from drowning, you have to first notice that she needs saving. Allocating your attention well is straightforwardly instrumental to fulfilling such duties. But how we allocate our attention can matter in its own right as well, that is, non-instrumentally. For one

[3] We want to stress that as long as you save the child, you do the right thing, on our view. Our point is just that, in doing the right thing, you can still exhibit better or worse moral agency through your attention allocation. We'll also flag here that this section develops an idealized conception of judicious attention allocation for theoretic purposes. In Sect. 3.5 we'll address the worry that our view is overly idealized.

thing, fulfilling our duties of attention can be constitutive of our morally significant roles and relationships. This is most apparent where further consequences are held constant.

If you're babysitting and you spend the whole time staring at your phone, paying no attention to the child in your charge, this makes you a negligent caregiver. If it turns out that she tucked herself into bed and went to sleep as soon as her parents left, then your outward behavior and its consequences may be no different than if you were paying attention to her. The moral difference, then, between negligence and responsibility here is evidently just a difference in attention allocation. As you pass the time on your phone, do you dutifully notice noises in the next room and listen for indications that everything is okay, or are you completely oblivious? Likewise, if you see that your best friend is depressed, but you know she'd rather not talk about it, you might pretend not to notice. If so, then you act no differently than if you actually failed to notice her poor state. But part of what it is to be a good friend is to notice such things. And the special attention friends owe to each other in virtue of their friendship doesn't just promote further goods for them—though of course it can and often does do that—it also *instantiates* the value they place on their relationship and *constitutes* their love for each other (Kolodny, 2003).

Duties of attention arguably obtain in less intimate relationships as well, such as the relationship between fellow citizens of a democratic polity. On January 6th 2021, US Americans plausibly owed it to each other to pay appropriate attention to the insurrection at the Capitol. If so, then those who indifferently scrolled past the headlines failed in their civic duties of attention, as did those who closely followed the developments with glee, hoping that the election results wouldn't be certified. Again, judicious attention allocation matters in such cases in part because of its effects on actions and outcomes. But it also matters in its own right. Suppose the insurrection had succeeded, suspending the mechanisms of self-government. Though this would have prevented democratic actions and outcomes, people qua democratic citizens could still have paid attention to these developments in morally better or worse ways. Dutifully attending to the collapse of the democratic institutions they hold dear would be constitutive of the bonds between them, instantiating their value for collective self-governance.[4]

Now, if we have duties of attention, and if we increasingly rely on automated systems to allocate our attention, this raises the question of whether these systems are properly aligned with our duties. It also raises a related question about whether outsourcing our practice of attention allocation to automated systems is itself morally problematic. We propose that judicious attention allocation is a moral skill. And

[4] Of course, others may have different intuitions about our examples here and about the non-instrumental value of attention allocation in general. If these differences come down to intuitional bedrock or fundamental differences in starting points for normative theory, we can't expect to convince our opponents. Even those who disagree with us on this point, however, can endorse the framework for moral assessment of attention allocation we propose, since the question of instrumentality is separate from the questions about first-order assessment on which we focus. Those who are inclined to a purely instrumental view of attention allocation can also agree with most of the concerns we raise about algorithmic recommendation in Sect. 3.

so, if reliance on automated systems deprives us of opportunities to develop and exercise this skill, it may pose a moral problem even if these systems were otherwise assistive for discharging our duties of attention. We address both questions in Sect. 2. But first, we complete the groundwork for our critique by developing a conception of attention as a moral skill.[5]

2.2 Attentional moral skill

For our purposes, understanding judicious attention allocation in terms of skill emphasizes three important points. Like skills in general, it (1) has success conditions, (2) develops through practice guided by those success conditions, and (3) is shaped by the tools used in its exercise. For comparison, consider the skill of a bicycle mechanic. This skill is conducive to repairing and maintaining bikes such that they ride swiftly, safely, and comfortably. These are among the normative standards against which mechanical skill is measured. And practice, guided by these standards, makes bike mechanics appropriately responsive to relevant factors, such as component alignment and cable tension. Skilled practitioners can even perceive such things differently than their unskilled counterparts (Stokes, 2021): where the expert mechanic sees worn-out gear teeth liable to cause chain-skipping, the novice sees only interlocking metal parts. This skill is also shaped by the tools one uses and the bikes one works on, among other factors. The skill of an amateur mechanic who restores garage-sale bikes with a basic toolset is markedly different from that of a professional who works on high-end racing bikes in a state-of-the-art shop.

Judicious attention allocation can helpfully be understood in essentially the same terms. First, as we've argued, it too is subject to normative standards. While the success conditions for bike maintenance and repair may be somewhat less contentious and more measurable than those for judicious attention allocation, this difference shouldn't be overstated. Standards like speed, safety, and comfort can demand very different things for different kinds of bikes and riders. Consider speed for a cargo bike versus a racing bike, safety for a child's training bike versus a BMX stunt bike, and comfort for a commuter bike versus a mountain bike. Similarly, the particular moral standards for allocating attention can vary depending on factors like the relationship (close friends versus strangers), the context (personal versus professional),

[5] The concept of moral skill is used in different ways by ethical theorists. Some treat it as synonymous with moral virtue (Annas, 2011); others treat it as a useful but limited analogue for virtue (Stichter, 2018); and still others see moral skill as a scaffolding which is necessary but not sufficient for full virtue (Vallor, 2015). Some think that moral skill is essentially the same kind of achievement as other skills (Swartwood, 2013), while others argue that moral skill is extremely difficult, if not impossible, to acquire (Shepherd, 2022). Some emphasize the deliberative, intellectual aspects of moral skill (Bloomfield, 2000), and others emphasize its intuitive, automatic aspects (Fridland, 2017). And some think the normative standards for moral skill are indexed to the individual's conception of the good life (Tsai, 2020), whereas others think the standards are agent-independent (Schuster, 2023). For present purposes, we bracket these debates and conceive of attention as a moral skill in the minimal sense we outline in Sect. 2.2.

and the broader cultural environment (the particulars here can be especially varied). If there is a difference, then, it's apparently one of degree rather than kind.

In any case, moral standards for attention allocation are generally clear enough to guide skill development. If your best friend suddenly becomes deeply depressed, this is obviously something you should pay attention to. Your more particular duties of attention will depend on various factors—how your friend exhibits depression, the cause of her depression, what's likely to help her, and so forth—but these are just the kinds of particulars good friends are typically aware of. Because they pay special attention to each other, over time they learn about each other's lives, values, habits, needs, and other things that enable them to tell when their attention is respectful versus intrusive, supportive versus smothering, caring versus self-interested, and so on. If this weren't the case, it would be hard to explain how good friends learn to attend well to each other's many particularities. Surely, they're not simply born with this ability or necessarily develop it regardless of how they interact with each other.

This highlights the second way in which judicious attention allocation is skillful: it requires practice (Annas, 2011). Bike mechanics aren't just born with the ability to tune derailleurs. They have to learn such things through trial and error, measuring their performance against the relevant standards (for example, the chain shouldn't rub against other components). Likewise, moral agents have to learn what to attend to, how so, how much, and for what reasons. This is especially clear when we consider how factors like culture, context, and specific relationships define duties of attention. A morally inexperienced child might stare in fixation at a person using a wheelchair, whereas a morally skilled adult will attend to this person in a way that acknowledges her particular needs without defining her in terms of them (Stohr, 2018). The difference is that the adult has had various forms of social and emotional feedback (Jacobson, 2005)—explicit instruction like "don't stare" as well as implicit guidance like the appreciative or irritated demeanor with which people respond to her attention—to teach her how to attend respectfully, whereas the child hasn't.

Even among adults, though, some have much more opportunity than others to learn how to attend appropriately to people with special needs, like those who have a family member with a wheelchair. So environmental factors can play a significant role in defining the particular shape of moral skills.[6] And this points to the third way in which judicious attention allocation is skillful: among the most important environmental factors for the development and exercise of skill are the relevant tools available. A bike mechanic can remove chains efficiently with a specialized chain-breaking tool, which enhances her mechanical skill while also shaping it in a particular way. A professional mechanic will be handy with a chain breaker but likely not with a hammer and punch (a cheap method used by some amateurs). Certain tools can have a similar influence on attentional moral skill. Distant friends might rely on phones, email, and social media to keep in touch, and this can shape how they fulfill their mutual duties of attention. Without much face-to-face interaction,

[6] For this reason we agree with Shepherd (2022) that moral skills can be highly context-dependent. But since many skills are similarly brittle (Kilov 2020), we don't think this counts against conceiving of judicious attention allocation as a moral skill.

they'll have to be good at interpreting each other's moods from things like tone of voice, content of messages, and frequency of posts, whereas noticing each other's subtle physical cues will be less important.

2.3 Deskilling

The tools on which we rely can also undermine skill, though: that is, they can cause deskilling. If you never learned to drive a car with a manual transmission or haven't done so for years, reliance on an automatic transmission has probably made you relatively bad at driving a manual. If you're hauling a heavy load or driving up a mountain with your automatic, then, you might well fail to manually engage the car's rarely used low-gear setting, leading to engine damage or worse. Similarly, if you routinely rely on Google Maps to find your way from A to B, you're more likely to make an unplanned detour to C when your phone battery dies.

Others have already argued that digital technologies in general can have a deskilling effect on attention allocation. Shannon Vallor (2015) contends that certain technologies can both misdirect attention and degrade attentional skill over time. Specifically, she claims that media multitasking—that is, "the practice of consuming multiple information streams at once (social media feeds, text messages, photos, internet videos, television, homework, music, phone calls, reading, etc.), and/or carrying on multiple information exchanges simultaneously" (p. 117)—can make it harder to focus on things that really matter. And over time, chronic media multitasking can evidently lead to diminished skill at focusing in general.

For instance, as Vallor notes, Ophir et al. (2009) find that people who routinely engage in heavy media multitasking are generally more susceptible to distraction than those who don't. And Wang and Tchernev (2012) point out that this undermines the cognitive needs that drive the practice in the first place (information, knowledge, and understanding) while satisfying emotional needs (habitual media use) in a way that's structurally similar to classical conditioning. So media multitasking not only makes it hard to focus on things that matter while doing it, it's also superficially rewarding in a way that encourages people to do it more and more. And in chronic cases, this practice is associated with attentional deficiencies.

This case exemplifies a more general point: technologies aren't just neutral means for pursuing whatever ends we choose; rather, their design influences our actions and aims in distinctive ways and thereby shapes us as agents (Verbeek, 2011). They enable certain actions while preventing others and encourage some modes of engagement while discouraging others. Such affordances don't deterministically force us to behave one way rather than another, but they do significantly constrain and influence our actions (Davis, 2020). In the case just discussed, certain digital technologies encourage multitasking and discourage sustained focus, making it difficult for people to pay attention to the right things, in the right ways, to the right degrees, and for the right reasons. No surprise, then, that over time these tools can evidently diminish attentional moral skill.

Consider now an additional set of concerns: while digital media technologies can encourage distraction, they can also encourage sustained focus. Recommender systems, in particular, are designed to capture and hold our attention. This is straightforwardly bad if these tools misallocate our attention. But algorithmic recommendation could be morally problematic even if it reliably aligned with our duties of attention. Just as outsourcing certain tasks to an automatic transmission can cause important driving skills to atrophy, or not develop in the first place, relying on automated systems to allocate our attention can plausibly lead to moral deskilling. And if we attend as we ought to only because automated systems direct us to do so, then even if we get things right in other respects we arguably don't do so for the right reasons.

Whether and to what extent algorithmic recommendation does in fact pose such a threat, however, depends on certain empirical and technical considerations. How much do we rely on algorithmic recommendation? Do we have reason to expect algorithmic recommendations to align with our duties of attention? And is reliance on recommender systems likely to make us generally worse at judiciously allocating our attention? We now turn to these questions.

3 Algorithmic recommendation and moral deskilling

3.1 Reliance on algorithmic recommendation

We won't attempt a comprehensive review of research on algorithmic recommendation. Our argument is conditional on the empirical hypothesis that the inhabitants of digital societies are already, or are likely soon to become, substantially dependent on recommendation algorithms to allocate their attention online. And this hypothesis is well supported by even a cursory overview of the relevant empirical literature.[7]

By recent estimates, nearly two thirds of people globally are now online, over 90 percent of whom are regular social media users. These figures have risen steadily with the proliferation of smartphones, which currently account for about half of all connected time in high-tech economies and far more in developing economies. The average internet user is online for over six and a half hours per day, including about two and half hours on social media. Reading articles and streaming videos, music, and podcasts on other popular platforms also collectively account for several hours per day for the average internet user.[8]

The content on major platforms, moreover, is now standardly filtered and ranked by recommender systems. In recent years, one platform after another has departed from subscription-based approaches to content distribution in favor of algorithmically-mediated approaches. So instead of just showing users posts from accounts

[7] Fuller treatment would go into greater detail about how recommendation algorithms interact with the design of online platforms, as well as other factors such as content moderation, to determine the content that we're exposed to online. Algorithms don't operate in isolation; they always depend on these other factors to have their effects. For further discussion, see Bucher (2018), Lazar (2023), Narayanan (2023). For our purposes, though, just the fact that our attention is being extensively directed by external forces is what matters most.

[8] These figures come from Kemp (2023).

they choose to follow, for example, platforms now standardly expose users to whatever content they're likely to engage with. While this has at times caused dismay among users, algorithmic recommendation has consistently prevailed, and trends indicate that this is unlikely to change (Narayanan, 2023). This means that not only are most people spending much of their day engaging with online content, often via devices they carry around with them everywhere they go, their online experience is also extensively curated by algorithms.

In principle, this could enhance our attentional moral skill. The amount of content on the internet is functionally infinite. None of us will more than scratch the surface in our lifetime. Powerful AI tools that filter and rank content into tractable, customized feeds could enable us to cut through this infoglut and allocate our attention to the right objects, and perhaps even to the right degrees, in the right ways, and for the right reasons. We explore this idea further in Sect. 4. But as recommender systems are currently designed and integrated into platforms, and our technosocial world more broadly, they're prone to allocating attention in morally problematic ways.

3.2 Automated attention (mis)allocation

One fundamental problem with algorithmic recommendation, in its current form, is the disconnect between our interests as moral agents and economic incentives for online platforms.[9] Algorithmic recommendation serves various business interests: keeping users coming back to platforms, directing them to products they're likely to buy, and exposing them to ads that generate revenue, to name a few. For such purposes, more engagement with platforms is better than less. Moreover, engagement can itself be monetized. Platforms collect vast amounts of data about users via their engagement, which they can then sell to other companies, generating additional revenue. Platforms therefore have strong incentives to design and implement their recommender systems to optimize for engagement.

While the business value of algorithmic recommendation is hard to measure precisely (Jannach & Jugovac, 2019), it's evidently effective enough to warrant the significant financial investment that platforms continually make in their recommender systems (Adomavicius et al., 2017; Gomez-Uribe & Hunt, 2015; Kumar & Hosanagar, 2019). And insofar as the incentives for platforms to maximize engagement are misaligned with users' duties of attention, algorithmic recommendation is apt to direct their attention in morally problematic ways. The four dimensions for moral assessment of attention allocation which we proposed in Sect. 2 help to specify these problems. Here we discuss objects and degrees of attention before turning to reasons for attending and ways of attending in 3.3 and 3.4.

[9] Indeed, our interests as moral agents can come apart from prevailing economic incentives quite generally, offline as well as online. This points to a likely objection to our project: there's nothing especially new or concerning about the problems we raise in this section. We explicitly address that objection in Sect. 3.5 below.

To begin with, algorithmic recommendation is evidently prone to directing our attention to things that don't matter to us, morally speaking, and even things that are positively morally bad for us to pay attention to. While it's in the interest of platforms not to expose their users to morally problematic content which could drive them away, typical moderation practices still allow the circulation of plenty of content people would be better off not paying attention to, such as the vast amount of misinformation about the legitimacy of the 2020 US election. In addition to encouraging violent behavior, engaging with lots of content like this can be bad in itself. Attending to items that reinforce one's identity as an election denier is central to being an election denier. And so, supposing that one ought not be an election denier, attending to such content is morally bad even if it doesn't lead one to violence.

But even for morally innocuous content, algorithmic recommendation tends to filter and rank things in morally problematic ways. For instance, Mansoury et al. (2020) demonstrate how recommender systems are prone to feedback loops that amplify popularity bias and increase homogenization of users' experiences. When certain content gets a lot of engagement, it tends to get recommended to many users. And the more users it gets recommended to, the more engagement it's likely to get, leading to even more widespread recommendation. Thus, huge numbers of users can end up being exposed to the same content regardless of the diversity of their antecedent interests. This can be bad because, for one thing, popular yet trivial content can crowd out content of greater moral significance. And even if popular content happens to be morally relevant to many users, it may have little to do with other users' particular duties of attention. The root problem here is that recommender systems optimize for engagement, not moral relevance. So even when morally important content happens to go viral, as ALS awareness did in 2014, this is merely coincidental. Just consider all the worthy causes that haven't gone viral and all the trivial and morally noxious content that has.

For the same reason, recommendation feedback loops can encourage excessive and deficient modes of attention, even when attention gets directed to appropriate objects. In addition to a lot of junk, the internet contains a vast amount of content worth attending to. To optimize for engagement, though, recommender systems steer each of us toward whatever we're most likely to engage with, even if we ought to pay more attention to other things. And when we engage with recommended content, this reinforces their tendency to prioritize similar content for us, which in turn encourages us to engage more with it. This can be content that's generally popular, but it can also be more idiosyncratic. Because of your viewing history, your YouTube feed might include lots of popular videos about a certain former president's legal troubles as well as many clips of your favorite podcast host's topical rants.

Some researchers argue that this sort of feedback loop creates echo chambers—"the effect of a user's interest being positively or negatively reinforced by repeated exposure to a certain item or category of items"—and filter bubbles, which occur when "recommender systems select limited content to serve users online" (Jiang et al., 2019, p. 383). While the evidence for these particular phenomena is contentious (Ross Arguedas et al., 2022), we need not assume that they're widespread, or that they exist at all, for present purposes. It only needs to be the case that recommender systems are prone to directing individual users' attention excessively to some things and deficiently to others.

It's important to keep in mind, then, that the whole purpose of algorithmic recommendation, from the standpoint of the platforms that rely on it, is to influence users to pay more attention to things that promote their engagement than to things that don't. If what we find most engaging isn't the same as what we ought to pay most attention to, then, recommender systems are apt to misallocate our attention in this respect. So even if we had no empirical evidence about how algorithmic recommendation affects attention allocation—say, if the technology existed but hadn't yet been deployed—we would still have reason to be concerned.

Considering how recommender systems are designed and implemented offers further evidence of why they're prone to allocating attention in morally problematic ways. Since these tools are meant to dynamically model and predict users' preferences based on their behavior, reinforcement learning (RL) is an especially powerful technique for training them.[10] Another common application of RL, which helpfully illustrates how it works, is game playing. RL can teach an AI system to play a game by specifying the object of the game in a reward function which reinforces goal-approximating behavior. With a well-specified reward function and enough opportunities to refine its approach through interaction with its environment, an AI system can learn to play a wide range of games, often at superhuman levels. But a poorly specified reward function can lead to reward hacking.

This happens when the AI system finds a way to satisfy the reward function without doing what the designers intend. OpenAI discusses its own failure to train an AI system to play CoastRunners, a boat racing video game, with RL (Clark & Amodei, 2016). They specified the reward function in terms of scoring points, a proxy for performing well in a race. But because players can score points by running into certain objects during the race, the AI learned to steer its boat in circles, repeatedly crashing into things and never finishing. Even so, this strategy enabled it to outscore most human players. While morally inconsequential in this case, reward hacking is seriously concerning for recommender systems designed to capture our attention. Rather than learning what we care about, these systems can instead learn to manipulate our preferences (Albanie et al., 2017; Krueger et al., 2020) in order to make them easier to predict and satisfy (Chaney et al., 2017), especially if they get rewarded for maximizing long-term engagement (Kasirzadeh & Evans, 2021).[11]

Moreover, both preferences and engagement get specified in problematic ways for recommender systems. The behavioral data they train on are at best coarse-grained proxies for what we really care about—things like our clicks, likes, ratings, and viewing times—and this is also the kind of engagement they aim to encourage. But such behaviors don't necessarily reflect our actual preferences, much less align with

[10] See Zerilli (2021) for a helpful overview of this and other common ML techniques. See Afsar et al. (2022) for details on RL techniques for recommender systems.

[11] Noting this danger, researchers have identified reward hacking as one of the most pressing problems in AI safety (Amodei et al., 2016), called for interdisciplinary research efforts to address it (Franklin et al., 2022), and emphasized the critical, even existential, importance of accurately specifying what AI systems should optimize for (Russell, 2020).

our duties of attention. In fact, how we tend to engage with content can be directly contrary to both.

You might be susceptible to clickbait with titles like "she was famous in the 90s, see what she looks like now" or "presidential announcement sparks 'total outrage'." But you might well have a higher-order preference not to be so inclined, if you care about being productive, focusing on accurate and important information, and limiting screen time more than the instant gratification that too often drives your behavior. Likewise, you might care deeply about fulfilling your duties of attention; but if you're more likely to engage with clickbait than posts from your depressed friend about her recent divorce, recommender systems are designed to encourage you to do so. Even if they do happen to direct your attention to your friend's posts, then, it would once again be merely coincidental if they also promoted her posts to an appropriate degree.

3.3 Automation and reasons-responsiveness

Suppose, though, that algorithmic recommendation wasn't used to maximize platform engagement and so wasn't prone to these problematic feedback loops. Instead, imagine that recommender systems reliably directed our attention to the right things and in the right degrees. Would the moral skill of judicious attention allocation necessarily be well served? Would relying on automated systems, and not our own reasons-responsiveness, realize the same moral value? We don't think so. Attending to something only because it's been served up to you isn't as good as doing so for your own moral reasons. Even if judicious attention allocation had no more instrumental value than outsourcing attention allocation to automated systems, then, it would still be non-instrumentally better.

In the case of the drowning child, we argued that it would be morally suboptimal to pay attention only because you fear legal consequences. One reason for this is that your response to the situation lacks counterfactual robustness. If you pay attention only out of self-interest, then if it didn't serve your interests, you wouldn't do it. Likewise, if you pay attention to your depressed friend's updates only because they get algorithmically recommended to you, then you wouldn't pay attention otherwise. And if that's the case, your attention allocation fails to fulfill your special duties of friendship.[12]

Now, we don't want to overstate this point. We humans are cognitively and morally frail creatures who often fail to attend appropriately to the people and things most dear to us even under favorable circumstances. Nonetheless, the morally significant bonds we share with others demand *some* robustness. If you all but forget that your depressed friend exists when she disappears from your social media feeds, this

[12] This is importantly distinct from a case where you use algorithmically curated platforms to help you keep tabs on your friend *because* you care about her. In that case, if you didn't see any posts from her for a while, you would presumably take notice, since your attention allocation would ultimately be driven by your concern for her and not just by your algorithmic scaffolding. We discuss evidence in Sect. 3.4, however, which casts doubt on how easy it is to use online platforms to enhance moral agency like this.

is at least morally suboptimal. And the longer it goes on, the more you fail in your attentional duties of friendship. At some point, you're just not really her friend at all.

The problem here is that, even when you pay attention to your depressed friend, if you rely too much on algorithmic recommendation, you don't do so for your own moral reasons. Why, then, do you attend to her? As long as it's sufficiently likely to promote your engagement with online platforms, recommender systems will tend to direct your attention to her regardless of why you respond this way. Maybe you're infatuated with her. Maybe you despise her. Maybe you just find her amusing. You may have no idea why you pay attention to her. But as long as you engage, this serves the commercial interests of those who design and deploy the recommender systems. And suffice it to say, the commercial interests of platforms aren't the sort of reasons that constitute morally significant relationships like friendship.

Moreover, even to the extent that your own preferences factor in here, if those preferences are only incidentally aligned with your duties of attention, this too can be a problem. Suppose you read your depressed friend's posts because you're taken in by the soap opera of her life. You can't get enough of the drama. But you don't like this about yourself. You do genuinely care about your friend, and you wish you could focus more on her needs than the tantalizing details of her personal life. This sort of critical self-reflection suggests that, though you do attend to the right things here, perhaps even to an appropriate degree, you don't do so for what you take to be your own *moral* reasons. Instead, you do so because of the regrettable preferences you express through your patterns of engagement. And because this is what recommender systems are designed to optimize for, they encourage you to allocate your attention for morally bad reasons, even though your own preferences drive the recommendation process.

Finally, it's worth noting that particular algorithmic recommendations are difficult, if not impossible, to fully explain due to the opacity of recommender systems. So even though we know that commercial interest and our own behavioral patterns make a difference, we can't know exactly how these things, among many others, factor into the recommendations we get in any given instance. Platforms sometimes allow access to their recommendation algorithms, but the vast majority of users lack the technical knowledge to make sense of computer code and statistical functions. And even those who design these systems can't explain the paths from inputs to outputs in ways that amount to normative reasons for particular outputs. Like many of today's AI technologies, recommender systems use deep learning with artificial neural networks to discover highly complex patterns among users' behavior, features of content, and many other factors, which elude human analysis. Inspired by the structure of biological brains, artificial neural networks find multidimensional patterns in vast datasets by iteratively adjusting the weights between nodes arranged in multiple layers between inputs and outputs. The weights, or parameters, in a single system of this kind can total in the trillions, making them extremely powerful and flexible but also humanly impossible to interpret (Burrell, 2016).[13] Due to this opacity, then,

[13] While simpler, more interpretable systems like decision trees don't face this problem to the same degree, they're still prone to finding multiple redundant paths from inputs to outputs (Izza et al., 2022). So they, too, can make it difficult to determine which factors are ultimately relevant in any given case. Interpretable ML is a live area of research, and it might eventually deliver AI systems or methods for

existing recommender systems can't give us normative reasons for directing our attention as they do. And if they can't give us such reasons, we can't endorse those reasons and adopt them as our own.

3.4 Automation and moral deskilling

Importantly, though, none of this is to say that when we pay attention to algorithmically recommended content we're *necessarily* failing to allocate our attention for ourselves. Suppose you judge that algorithmic recommendations align well enough with your duties of attention most of the time, and you remain diligent about how you allocate your attention online. If so, then you could use this technology to aid, or at least not interfere with, judicious attention allocation. This seems perfectly possible. But certain affordances of recommender systems and the platforms that deploy them raise serious doubts about how easy it is to exercise executive oversight like this. We focus here on two: (1) like other automated systems, they encourage users to rely on them and make it hard to critically monitor their performance; and (2) they can co-opt and distort moral responsiveness itself. Together, these affordances not only pose a challenge for exercising the oversight necessary to use recommender systems morally well, they threaten to undermine attentional moral skill altogether.

Regarding (1), when a task that formerly required skill becomes effectively automated, people have less motivation and opportunity to learn to do it, and those who already have the necessary skill have less motivation and opportunity to maintain it. Further, not only does lack of practice tend to make people less skilled at tasks that get automated, it evidently also makes it hard for them to effectively monitor the automated performance of those tasks. This is because people tend to pay attention to a task only to the extent that the task demands it; and so when an automated task requires little attention, people tend to allocate little attention to it (Walker et al., 2015).

Semi-autonomous cars provide a vivid illustration of what can happen next: when these heavily automated machines perform sufficiently well, people tend to stop monitoring them diligently enough to notice failures and intervene, sometimes with lethal consequences (Cunningham & Regan, 2018). This is one of the "ironies of automation" Lisanne Bainbridge (1983) identifies in her seminal research on human factors: advanced automated systems often still need some human supervision, but they tend to diminish their human operators' capacity and propensity to exercise effective oversight.[14]

We see no reason to think that the effects of algorithmic recommendation on attentional moral skill should be any different. As Vallor points out, "moral skills appear just as vulnerable to disruption or devaluation by technology-driven shifts in human practices as are professional or artisanal skills such as machining,

Footnote 13 (continued)
interpreting them which give users straightforward, veridical reasons for their outputs (Rudin et al., 2022). But this is as yet an unfulfilled promise.

[14] Relatedly, see Zerilli et al. (2019) on the "control problem" posed by AI systems.

shoemaking, or gardening" (2015, p. 109). Relying on recommender systems to direct our attention deprives us of opportunities to practice allocating our attention through our own offices and for our own moral reasons. And it's very plausible that when these tools direct our attention well enough to keep us engaged, complacency creeps in.

The fact that these tools cut through the information saturation of the internet with an efficiency we could never hope to achieve without them, and that their inner workings are largely opaque to us, also encourages us to defer to them about what's worthy of our attention rather than critically monitor their outputs (Bainbridge, 1983, p. 776). It's much easier to just passively rely on our algorithmic scaffolding than to take an active, executive role here. And as many of us now spend much of our day engaging with algorithmically-curated online platforms, it's hard to imagine that the attentional complacency they encourage could have no deleterious effect on our ability to judiciously allocate our attention for ourselves, on those platforms as well as beyond them. Much like a sedentary job is bad for one's general physical fitness, it stands to reason that the "sedentary" attentional lifestyle afforded by algorithmic recommendation is bad for one's general attentional fitness.

Regarding (2), even if attentional moral skill turned out to be resilient to such affordances, algorithmic recommendation still poses a further challenge: recommender systems have an evident tendency to actively engage, and misdirect, our moral responsiveness. As a moral skill, judicious attention allocation involves responding appropriately to morally relevant factors such that we attend to them in the right ways, to the right degrees, and for the right reasons. But our moral responsiveness is also a major driver of how we allocate our attention simpliciter, making it a prime target for recommender systems to exploit. According to Brady et al. (2020), moralized digital content tends to get more engagement than morally neutral content, because it tends to incite stronger reactions; and so recommender systems have a built-in incentive to exploit our moral responsiveness.

This, in turn, incentivizes content creators to produce more moralized content in order to advance their own interests. Brady et al. (2020) therefore warn that "moral and emotional appeals that capture attention can be exploited by disinformation profiteers, as in the case of fake news spread around the 2016 U.S. election." But even for content that happens to be worthy of our attention, they explain that "people are motivated to share moral-emotional content based on their group identity" among other quasi-moral impulses that are prone to misfiring. And they argue that "the design of social-media platforms interacts with these psychological tendencies" to facilitate the viral spread of moralized content, an effect they call "moral contagion." This helps to explain the prevalence of heated, often toxic, moral discourse online. It's not just human nature, it's the affordances of algorithmically curated platforms bringing out the worst in human nature.

Again, though, these affordances don't *necessarily* encourage the wrong kind of attention allocation. Some things, like police brutality, sexual violence, and political corruption, should exogenously grab and hold our attention via strong moral emotions. The problem is that recommender systems are prone to encouraging this kind of response regardless of whether the content warrants it. Even issues that demand more careful, deliberate attention allocation are likely to get more engagement if they evoke strong moral-emotional responses. As the name suggests, moral "contagion" enables content to "go viral." And so recommender systems are apt to exploit our tendency to respond strongly to moralized content, including for morally problematic reasons like perceived threat to group identity.

In addition to encouraging complacency and discouraging judicious attention allocation, then, the affordances of recommender systems also threaten to (re)shape moral responsiveness in the service of maximizing platform engagement rather than in the service of attentional moral skill and, ultimately, the fulfillment of duties of attention. Once again, it's hard to imagine that the effects of spending so much time on platforms that have strong incentives to shape us in this way could be confined only to our behavior on those platforms. We should take seriously the possibility that they shape our attentional capacities and propensities more generally, in ways that transfer to other contexts, including offline ones.

3.5 Objections

Before turning to our proposal for addressing the concerns we've raised in this section, we'll address three likely objections to our critique: (1) we haven't provided sufficient evidence of a direct causal link between algorithmic recommendation and attentional moral deficits; (2) even granting that algorithmic recommendation threatens judicious attention allocation, it may not do so in an especially serious or novel way; and (3) our view of attentional moral skill is overly idealized, exaggerating the problems we've identified.[15]

Beginning with (1), we acknowledge that the empirical evidence we've adduced to support our critique is indirect and correlative. This reflects limitations on empirical investigation into the effects of algorithmic recommendation. Running simulations with recommender systems can reveal potential relationships among some variables; but this sort of evidence lacks ecological validity, since there are many factors in the world which simulations can't account for. Observing actual online behavior from the outside, by analyzing available platform data, can provide a more realistic picture; but this reveals primarily correlations, not causal relationships. And even with full access to conduct experiments on a major platform, important factors in the world beyond would be left out, including what's happening on other platforms.[16] Finally, at the time of writing, none of the major online platforms offer this kind of access to independent researchers. So it would be perverse to absolve them of responsibility for likely adverse individual and social impacts on the basis that we can't adequately demonstrate, from the outside, a causal relationship between their design decisions and those impacts. Our best hope of doing so is obstructed by their refusal to provide us with sufficient access to their data and systems.

It's also important to recognize that the same kinds of problems plague any inquiry into the effects of new technologies, operating in complex sociotechnical

[15] We thank two anonymous referees for pressing these objections.

[16] See Thornburn et al. (2022) for further discussion of these limitations.

environments, on human agency and wellbeing. But these challenges don't require us to suspend judgment altogether, just to qualify it. And if we're right that judicious attention allocation is an important moral skill, then it's crucial to determine to the best of our ability whether current technosocial conditions threaten it. We've made the affirmative case with the resources we have: the framework we've developed for moral assessment of attention allocation; our account of attention as a moral skill; available evidence of how algorithmic recommendation affects attention allocation; and evidence for the general deskilling effects of automated systems. At the very least, this provides grounds for further investigation and consideration of alternative methods for navigating the online world.

Objection (2) grants that algorithmic recommendation threatens attentional moral skill but questions the novelty and gravity of this threat. Certainly, many things have always competed for human attention. And powerful external forces have long had interests, economic and otherwise, in shaping our attentional capacities and propensities, potentially in morally problematic ways. For instance, businesses want our money, politicians want our votes, and they can be all too willing to employ manipulative tactics that work against our interests as moral agents. Through analogue tools like traditional mass media, they too can therefore allocate our attention to the wrong things, in the wrong ways, to the wrong degrees, and for the wrong reasons, thwarting and undermining our attentional moral skill. We acknowledge all this. But we contend that algorithmic recommendation accelerates and exacerbates such problems. Moreover, it can shape individual and collective attention in new ways, opening an especially problematic gap between attention allocation and the values that should drive it.

Without AI, it's possible to identify and exploit individual susceptibilities to influence attention, and it's also possible to exert influence on the collective attention of large populations, but it's very hard to do both at once. Close friends have intimate knowledge about each others' proclivities, whereas billboards have wide public reach, but few things in the offline, analogue world have both. The power of algorithmic recommendation lies in its ability to powerfully influence attention allocation through deeply personalized messaging coordinated to have maximal collective effect (Benn & Lazar, 2022).

Especially when driven by RL, recommender systems can learn our subtle, private behavioral tendencies and use this information to drive coordinated action. Human agents can sometimes harness this power for their own purposes, such as depressing voter turnout to get an unpopular candidate elected.[17] But algorithmic recommendation can have such population-level effects even without this sort of intentional guidance. Most viral trends take off more or less randomly. In any case, algorithmic recommendation's ability to combine deep individual personalization with large-scale collective influence makes it an especially powerful force acting on attention allocation. And the fact that internet users are so frequently exposed to its

[17] Donald Trump's 2016 presidential campaign, to take a prominent example, has been criticized for using Facebook's algorithmic infrastructure to microtarget highly customized campaign ads to users in order to influence voter turnout on a national scale. See Stahl (2017).

influence makes it especially threatening to the development, exercise, and maintenance of attentional moral skill.

This takes us to objection (3), which questions whether our view of attentional moral skill is overly idealized to begin with and exaggerates the problems we've identified. With or without algorithmic recommendation, people's attentional capacities and propensities are undoubtedly shaped by all kinds of potentially problematic factors, many of which they're not consciously aware of. How realistic is it, then, to expect them to attend to the right things, in the right ways, to the right degrees, and for the right reasons, even under favorable circumstances? We acknowledge that judicious attention allocation, like other practical skills, can be hard to develop and exercise. It may not be realistic, then, to expect people to allocate their attention reliably well much beyond the particular circumstances of their day-to-day lives, where they have the opportunity to learn through repeated practice. But even if attentional moral skill is brittle in this way, it can still be instrumental to acting well, constitute morally valuable relationships, and instantiate good moral values as far as it goes. You may be excused for not attending well to the poor children asking you for money while you try to navigate the train system in a foreign country. But if you have children of your own, you should at least be able to attend to their needs reliably well. And judicious attention allocation is well worth striving for even if we can only hope to achieve it in a limited, context-specific capacity.

Our argument, then, isn't that algorithmic recommendation prevents people from being perfect moral agents, free to develop and exercise attentional moral skill in its purest and best form. Rather, our view is that, in a world that already poses serious challenges for judicious attention allocation, algorithmic recommendation adds yet another challenge, and an especially powerful one at that. But it doesn't have to. The problems we've identified aren't intractable. We'll close, then, by considering how algorithmic recommendation might be reimagined such that it supports, rather than undermines, attentional moral skill.

4 Technomoral upskilling

Faced with the crushingly efficient machinery of information capitalism, it's easy to feel pessimistic. The drive to misallocate our attention sustains the most powerful and highly capitalized private companies the world has yet seen. The technology powering algorithmic recommendation, RL, is inherently prone to reward hacking: in this case, manipulating our behavior. And recommender systems rely on vast amounts of behavioral data, meaning that only the largest platforms have the resources to develop consequential tools of this kind. New regulations have recently been passed in the EU, in particular the Digital Services Act,[18] which could in theory support a better approach to algorithmic recommendation. But implementation and enforcement promise to be an uphill battle, fought tooth and nail by the major platforms at every step.

[18] Details are available at: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.

In response to this predicament, some researchers have argued that we should build recommender systems that can be layered on top of the existing platform ecosystem, as a kind of "middleware" (Fukuyama, 2021). The idea is that personal recommender systems could filter and rank the content on platforms without themselves being controlled by those platforms. While this is an attractive proposal, little progress has yet been made, for some fairly fundamental reasons. Existing recommender systems rely on ML techniques, like deep learning, which require vast amounts of behavioral data to function. And since such middleware would compete with platforms' own recommender systems, platforms have no incentive to share their data (Keller, 2021).[19] But even if they did, as long as these intermediaries still depended on RL, they would continue to face the problems we've identified with this approach to algorithmic recommendation. As long as the fundamental business model of information capitalism remains intact, it's hard to see how such a surface change could make a significant difference.

Recent advances in large language models (LLMs), however, suggest the possibility of developing a truly independent kind of middleware that could, in principle, reshape the attention economy and scaffold attentional moral skill rather than bypassing, co-opting, or otherwise undermining it.[20] LLMs are highly complex mathematical models of language, pre-trained using self-supervised deep learning on vast corpuses of text to predict missing words (or, more precisely, tokens that are either words or parts of words). These pre-trained models are then fine-tuned using supervised and reinforcement learning to identify not just the most likely completion of a given prompt but the one that's most engaging, factual, even helpful, harmless, and honest.[21]

While only specifically trained to predict missing tokens, these models have shown remarkable capabilities to perform many other functions. In addition to fluently generating and classifying natural language, the most powerful among them can be trained to use other software tools by calling on Application Programming Interfaces (APIs) to perform functions they can't perform on their own (Schick et al., 2023). For example, OpenAI's GPT-4 can use a range of plugins to browse the web and consult a symbolic AI system to answer mathematical questions. Complex systems like these, in which an LLM functions as the executive control center of a multifaceted system, are *generative agents*. In theory, it should be relatively simple to design a generative agent whose function is to browse platforms on your behalf—a kind of digital content sommelier or attention guardian (Friedman et al., 2023).

Online attention allocators have three basic tasks: determine the user's values; classify the content they could be exposed to; and predict which content will map onto the user's values. RL agents perform these tasks by leveraging massive amounts of behavioral data to infer people's preferences. By contrast, while building a generative agent also requires massive datasets, these systems train on text (and in

[19] Keller also argues that sharing such data would also have significant data protection implications.

[20] This proposal draws on Lazar (Forthcoming). See also Lazar (2023).

[21] See Lubbad (2023) for a helpful overview of OpenAI's model, GPT-4. See Bai et al. (2022) for how such models might be aligned with certain moral values.

some cases images) rather than behavior in order to learn how to engage with users conversationally. This doesn't amount, then, to the kind of mass surveillance that drives RL. Instead, generative agents build a sophisticated model of the world via our representations of it. And while these systems remain unlike human minds in many ways, they're well-suited to augment our mental powers.

For the purposes of allocating our attention, generative agents could (functionally) understand our moral values by asking us about them, including follow-up questions to clarify our answers. They could refine their models further by discussing particular recommendations with us. They could discuss our higher- and lower-order preferences with us, and we could instruct them to nudge us when we slip too much into satisfying the latter at the expense of the former. And they could even fill in the gaps around our stated preferences, as they already do in other conversational contexts. Indeed, this is one key to what makes them so effective as dialogue agents: they're specifically trained to be helpful, which requires them to account for things like the intent behind a statement.

Their facility with natural language, and in some cases with images as well, also means that generative agents should be adept at understanding the content of online communications. Moreover, they aren't confined to assessing discrete pieces of content in isolation but can set them in the context in which they appear (some models can process as many as 50,000 words of context at once). Given the ability to represent digital content in this rich way and to elicit and understand our values in similarly rich terms—not just the superficial desires and impulses revealed by our behavioral patterns, but what we genuinely care about most—in addition to their demonstrated reasoning capabilities, generative agents should be able to match content to values.

Rather than only having the crude signal of a thumbs-up or thumbs-down, then, you could train your own recommender system using natural language. Rather than playing the one-sided game of maximizing platform engagement, this sort of recommender system could play a two-sided game with you, for instance by asking you to rank lists of items in order to better understand what you mean when you say things like "I don't want to see content that pisses me off unless it's really important." With a natural language description of your values, as well as of the content it's recommending to you, a generative agent such as this should even be able to give you explanations for particular recommendations that amount to normative reasons: for instance, "posts about the culture war tend to make you angry, but this one is about an important upcoming referendum."[22]

In addition to providing a novel approach to filtering and curating online content, generative agents like these could also profoundly change the digital attention economy. Because they wouldn't rely on behavioral data collected and controlled

[22] Though AI systems still don't plausibly take normative reasons *as such*, at least not in the same way humans do (Véliz, 2021), their ability to articulate such reasons would enable generative agents to functionally enter the space of reasons in a way that current recommender systems can't (Heinrichs & Knell, 2021). This isn't to say, of course, that the reasons they would give us would necessarily be good. But when they fail to give us good reason, we could reject their recommendations on those grounds.

by platforms, they would shift power toward users. They would most plausibly be funded through subscription and could be integrated into operating systems, which have less incentive to maximize our engagement or to allocate our attention to any particular platform or type of content. Without such problematic incentives, generative agents could then prioritize facilitating our judicious attention allocation, responding directly to our stated moral values and thereby augmenting our attentional moral skill. Since they wouldn't be driven by RL, they wouldn't be prone to reward hacking through preference manipulation. And since such systems can interact with us conversationally, nor would they be inscrutable (at the relevant level), as existing recommender systems are. Instead, they could engage us in the recommendation process on the level of normative reasons. And this, in turn, would encourage us to maintain an active role in allocating our attention in the digital world.

Such systems no doubt face significant hurdles. For instance, they're currently vulnerable to attacks like prompt injection, which circumvent their training and subvert their purposes to those of the hacker (Greshake et al., 2023). And if hijacked, they could be more manipulative even than RL systems, due to their creativity and facility with natural language. There are also open questions about whether users would in practice be willing to invest the time needed to train up their own attention guardians. There are further, more technical challenges too (Friedman et al., 2023). If they can be resolved, however, generative agents offer perhaps the most exciting prospect yet of a new technological paradigm for AI-assisted attention allocation.

5 Conclusion

Major advances in information technology often cause moral panic about how they'll monopolize our attention, divert us from what really matters, and make us worse moral agents. We need to weigh such concerns with caution, as the causal connections between new technologies and social behaviors are always complex and contestable. Yet we can recognize general trends, such as that automating skillful tasks typically undermines the acquisition and retention of the relevant skills and that economic incentives often don't align with moral duties. Of course, nothing here is predetermined. Incentives and affordances can be counteracted through conscientious adaptation. But we do have to be conscientious in counteracting them.

Our existing approach to online attention allocation is importantly misaligned with our duties of attention and can evidently interfere with our development and maintenance of attentional moral skill. This shouldn't surprise us, since the systems that direct our attention are optimized for the maximization of returns on investment for the private companies that dominate the internet. To counteract these trends, we need to design systems that can augment, scaffold, and nurture judicious attention allocation rather than circumventing or undermining it. Dependence on reinforcement learning and big data for algorithmic recommendation makes such systems seem a forlorn hope. But new developments in generative AI offer the tantalizing prospect of a new kind of recommender system, one that could make us better, rather than worse, moral agents.

Author's contribution Both authors made substantial contributions to this project at every stage.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This project was funded in part by the Templeton World Charity Foundation.

Data availability No new data were created or analyzed for the purpose of this research.

Declarations

Conflict of interest The authors have no competing interests to report.

Ethical approval No ethics approval was required for this research.

References

Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2017). Effects of online recommendations on consumers' willingness to pay. *Information Systems Research, 29*(1), 84–102.

Albanie, S., Shakespeare, H., & Gunter, T. (2017). Unknowable manipulators: Social network curator algorithms. abs/1701.04895.

Allport, A. (1987). Selection for action: Some behavioural and neurophysiological considerations of attention and action. In H. Heuer & A. F. Sanders (Eds.), *Perspectives on perception and action* (pp. 395–419). Lawrence Erlbaum Associates.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv:1606.06565.

Annas, J. (2011). *Intelligent virtue*. Oxford University Press.

Afsar, M. M., Crump, T., & Far, B. (2022). Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys, 55*(7), 1–38.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, S., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., El Showk, S., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., & Kaplan, J. (2022). Constitutional AI: Hamlessness from AI Feedback. abs/2212.08073.

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19*(6), 775–779.

Benn, C., & Lazar, S. (2022). What's wrong with automated influence. *Canadian Journal of Philosophy, 52*(1), 125–148.

Berkman, E. T., Hutcherson, C. A., Livingston, J. L., Kahn, L. E., & Inzlicht, M. (2017). Self-control as value-based choice. *Current Directions in Psychological Science, 26*, 422–428.

Bermúdez, J. P. (2017). Social media and self-control: The vices and virtues of attention. In C. G. Prado (Ed.), *Social media and your brain: Web-based communication is changing how we think and express ourselves* (pp. 57–74). Praeger.

Bhargava, V. R., & Velasquez, M. (2021). Ethics of the attention economy: The problem of social media addiction. *Business Ethics Quarterly, 31*(3), 321–359.

Bloomfield, P. (2000). Virtue epistemology and the epistemology of virtue. *Philosophy and Phenomenological Research, 60*(1), 23–43.

Bommarito, N. (2013). Modesty as a virtue of attention. *The Philosophical Review, 122*(1), 93–117.

Brady, M. S. (2013). *Emotional insight: The epistemic role of emotional experience*. Oxford University Press.

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science, 15*(4), 978–1010.

Bucher, T. (2018). If...Then: Algorithmic Power and Politics. *Oxford Studies in Digital Politics*.

Bueno, C. C. (2016). *The attention economy: Labour, time and power in cognitive capitalism*. Rowman & Littlefield International.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 2053951715622512.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research, 51*(13), 1484–1525.

Chaney, A. J. B., Stewart, B. M., & Engelhardt, B. E. (2017) How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM conference on recommender systems*. 224–232.

Chappell, R. Y., & Yetter-Chappell, H. (2016). Virtue and salience. *Australasian Journal of Philosophy, 94*(3), 449–463.

Clark, J., & Amodei, D. (2016). Faulty reward functions in the wild. *OpenAI*. https://openai.com/research/faulty-reward-functions

Cunningham, M., & Regan, M. (2018). Automated vehicles may encourage a new breed of distracted drivers. *The Conversation*. https://theconversation.com/automated-vehicles-may-encourage-a-new-breed-of-distracted-drivers-101178

Davis, J. L. (2020). *How artifacts afford: The power and politics of everyday things*. MIT Press.

De Sousa, R. (1990). *The rationality of emotion*. MIT Press.

Franklin, M., Ashton, H., Gorman, R., & Armstrong, S. (2022). Recognising the importance of preference change: A call for a coordinated multidisciplinary research effort in the age of AI. arXiv:2203.10525.

Fridland, E. (2017). Motor skill and moral virtue. *Royal Institute of Philosophy Supplement, 80*, 139–170.

Friedman, L., Ahuja, S., Allen, D., Tan, Z., Sidahmed, H., Long, C., Xie, J., Schubiner, G., Patel, A., Lara, H., Chu, B., Chen, Z., & Tiwari, M. (2023). Leveraging large language models in conversational recommender systems. abs/2305.07961.

Fukuyama, F. (2021). Making the internet safe for democracy. *Journal of Democracy, 32*(2), 37–44.

Gomez-Uribe, C. A., & Hunt, N. (2015). The netflix recommender system: algorithms, business value, and innovation. *ACM Transactions on Management Information Systems, 6*(4), 1–19.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. abs/2302.12173

Heinrichs, B., & Knell, S. (2021). Aliens in the space of reasons? On the interaction between humans and artificial intelligent agents. *Philosophy of Technology, 34*, 1569–1580.

Hindman, M. (2018). *The internet trap: How the digital economy builds monopolies and undermines democracy*. Princeton University Press.

Izza, Y., Ignatiev, A., & Marques-Silva, J. (2022). On tackling explanation redundancy in decision trees. *Journal of Artificial Intelligence Research, 75*, 261–321.

Jacobson, D. (2005). Seeing by feeling: Virtues, skills, and moral perception. *Ethical Theory and Moral Practice, 8*(4), 387–409.

Jannach, D., & Jugovac, M. (2019). Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems, 10*(4), 1–23.

Jiang, R., Chiappa, S., Lattimore, T., György, A., & Kohli, P. (2019). Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 383–390.

Kasirzadeh, A., & Evans, C. (2021). User tampering in reinforcement learning recommender systems. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*.

Keller, D. (2021). The future of platform power: Making middleware work. *Journal of Democracy, 32*(3), 168–172.

Kemp, S. (2023). Digital 2023: Global overview report. *Datareportal*. https://datareportal.com/reports/digital-2023-global-overview-report

Kilov, D. (2020). The brittleness of expertise and why it matters. *Synthese, 199*, 3431–3455.

Kolodny, N. (2003). Love as valuing a relationship. *The Philosophical Review, 112*(2), 135–189.

Krueger, D. S., Maharaj, T., & Leike, J. (2020). Hidden incentives for auto-induced distributional shift. arXiv:2009.09153.

Kumar, A., & Hosanagar, K. (2019). Measuring the value of recommendation links on product demand. *Information Systems Research, 30*(3), 819–838.

Lazar, S. (2023). Communicative justice and the distribution of attention. *Knight First Amendment Institute*. https://knightcolumbia.org/content/communicative-justice-and-the-distribution-of-attention

Lazar, S. (Forthcoming). *Connected by code: Algorithmic intermediaries and political philosophy*. Oxford University Press.

Lubbad, M. (2023). The ultimate guide to GPT-4 parameters: Everything you need to know about NLP's game-changer. *Medium*. https://medium.com/@mlubbad/the-ultimate-guide-to-gpt-4-parameters-everything-you-need-to-know-about-nlps-game-changer-109b8767855a

Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., & Burke, R. (2020). Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2145–2148.

Mole, C. (2021). Attention. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/win2021/entries/attention/

Narayanan, A. (2023). Understanding social media recommendation algorithms. *Knight First Amendment Institute*. https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms

Neumann, O. (1987). Beyond capacity: A functional view of attention. In H. Heuer & A. F. Sanders (Eds.), *Perspectives on perception and action* (pp. 361–394). Lawrence Erlbaum Associates.

Ophir, E., Nass, C., & Wagner, A. D. (2009). Cognitive control in media multitaskers. *Proceedings of the National Academy of Sciences, 106*(37), 15583–15587.

Orquin, J. L., & Loose, S. M. (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica, 144*(1), 190–206.

Ross Arguedas, A., Robertson, C., Fletcher, R., & Nielsen, R. (2022). *Echo chambers, filter bubbles, and polarisation: A literature review*. Reuters Institute for the Study of Journalism.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys, 16*, 1–85.

Russell, S. (2020). Artificial intelligence: A binary approach. In S. M. Liao (Ed.), *Ethics of artificial intelligence* (pp. 327–341). Oxford University Press.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. arXiv:2302.04761

Schuster, N. (2023). The skill model: A dilemma for virtue ethics. *Ethical Theory and Moral Practice, 26*(3), 447–461.

Shepherd, J. (2022). Practical structure and moral skill. *Philosophical Quarterly, 72*(3), 713–732.

Simon, H. A. (1971). Designing organizations for an information-rich world. In M. Greenberger (Ed.), *Computers, communication, and the public interest* (pp. 37–52). Johns Hopkins University Press.

Stahl, L. (2017). Facebook "embeds," Russia and the Trump campaign's secret weapon. *CBS News*. https://www.cbsnews.com/news/facebook-embeds-russia-and-the-trump-campaigns-secret-weapon/

Stichter, M. (2018). *The Skillfulness of Virtue: Improving Our Moral and Epistemic Lives*. Cambridge University Press.

Stohr, K. (2018). Pretending not to notice: Respect, attention, and disability. In A. Cureton & T. Hill (Eds.), *Disability in practice: Attitudes, policies, and relationships* (pp. 50–71). Oxford University Press.

Stokes, D. (2021). On perceptual expertise. *Mind and Language, 36*(2), 241–263.

Swartwood, J. D. (2013). Wisdom as an expert skill. *Ethical Theory and Moral Practice, 16*, 511–528.

Thornburn, L., Stray, J., & Bengani, P. (2022). How to measure the effects of recommenders. *Medium*. https://medium.com/understanding-recommenders/how-to-measure-the-causal-effects-of-recommenders-5e89b7363d57

Tsai, C. (2020). Phronesis and techne: The skill model of wisdom defended. *Australasian Journal of Philosophy, 98*(2), 234–247.

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology, 23*, 107–124.

Véliz, C. (2021). Moral zombies: Why algorithms are not moral agents. *AI and Society, 36*(2), 487–497.

Verbeek, P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press.

Walker, G. H., Stanton, N. A., & Salmon, P. M. (2015). *Human factors in automotive engineering and technology*. Ashgate.

Wang, Z., & Tchernev, J. M. (2012). The myth of media multitasking: reciprocal dynamics of media multitasking, personal needs, and gratifications. *Journal of Communication, 62*(3), 493–513.

Vance, J., & Werner, P. J. (2022). Attentional moral perception. *Journal of Moral Philosophy, 19*(5), 501–525.

Waggoner, M. (2021). The focus of virtue: Attention broadening in empirically informed accounts of virtue cultivation. *Philosophical Psychology, 34*(8), 1217–1245.

Watzl, S. (2017). *Structuring mind: The nature of attention and how it shapes consciousness*. Oxford University Press.

Watzl, S. (2022). The ethics of attention: An argument and a framework. In S. A. Archer (Ed.), *Salience: A philosophical inquiry.* Routledge.

Williams, J. (2018). *Stand out of our light: Freedom and resistance in the attention economy*. Cambridge University Press.

Wu, W. (2011). Confronting many-many problems: Attention and agentive control. *Noûs, 45*(1), 51–60.

Wu, W. (2016). Experts and deviants: The story of agentive control. *Philosophy and Phenomenological Research, 93*(1), 101–126.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Algorithmic decision-making and the control problem. *Minds & Machines, 29*, 555–578.

Zerilli, J. (2021). What is artificial intelligence? In J. Zerilli (Ed.), *A Citizen's guide to artificial intelligence* (pp. 1–20). MIT Press.