



Agential harms of chatbots: an ethical perspective

Shervin MirzaeiGhazi¹ · Magda Osman^{2,3}

Received: 10 January 2026 / Accepted: 22 June 2026 / Published online: 13 July 2026
© The Author(s) 2026

Abstract

The widespread integration of AI-assisted digital technologies, particularly those leveraging Large Language Models (LLMs) and generative AI in applications like chatbots, presents both significant benefits and novel ethical challenges. While these technologies enhance efficiency and democratise access to services, their rapid deployment necessitates a critical examination of potential harms. There are ambitious forecasts of the revenue the chatbot industry is expected to generate over a 5-year window, and given the accelerated deployment, this paper examines its ethical risks. We introduce the concept of “agential harms”, which subsumes many others (psychological, developmental, recognition) that are often referred to as intangibles. Unlike traditional harms (physical, financial), agential harms subvert a person’s capacity to be an agent—to self-govern, develop, and be socially recognised—caused by an AI system’s convincing imitation of agency. These harms occur when the AI, by acting as a conversational partner or cognitive substitute, subtly undermines the very faculties that define human agency. Using AI assistant chatbots as a test case, we clarify the nature and significance of agential harms, distinguishing them from conventional categories of harms.

Keywords AI companions · Chatbots · AI risks · Agential harms · Ethical human-AI interactions

1 Introduction

AI-assisted digital technologies are advanced computational systems that use artificial intelligence to enhance or automate various tasks and processes, often working in conjunction with human users. These technologies frequently use Large Language Models (LLMs) and generative AI. LLMs refer to deep learning models pre-trained on vast amounts of data that enable them to generate and translate natural language and perform tasks like summarisation and question answering [70]. By generative AI, we mean the use of AI to

create new content and ideas, including text, images, audio, and more, by learning patterns from existing data [49].

The disparity between the benefits of generative AI and its clinical and economic realities necessitates a critical re-evaluation of the risks these systems pose, especially given the enormous enthusiasm for exposing the public to them. Chatbot applications are a growth industry; they are estimated to have generated approximately \$1.2 billion in 2024, and this will increase to \$2.1 billion by 2034 for chatbot-based mental health apps specifically (According to Insight-Ace Analytic).¹ However, macroeconomic analyses, such as those conducted by Goldman Sachs, indicate that despite widespread corporate adoption, approximately 80% of companies report no measurable impact on productivity. Such data suggests that generative AI investments have added “basically zero” to U.S. economic growth in the preceding year.² Despite this, promoting generative AI to ensure high-level corporate and public adoption has not waned.

Similarly, the deployment of LLM-powered chatbots has been contentious because of scepticism over their value added in terms of operational efficiencies [2, 18, 126, 134].

✉ Shervin MirzaeiGhazi
shervin.mirzaeighazi@fil.lu.se

Magda Osman
m.osman@jbs.cam.ac.uk

¹ Department of Philosophy, Lund University, Lund, Sweden

² Judge Business School, University of Cambridge, Cambridge, United Kingdom

³ Centre for Decision Research, Analytics Technology and Operations Department, Leeds Business School, University of Leeds, Leeds, England, UK

¹ <https://www.insightaceanalytic.com/request-sample/1976>

² <https://gizmodo.com/ai-added-basically-zero-to-us-economic-growth-last-year-goldman-sachs-says-2000725380>

Further still, there are profound concerns when it comes to chatbots employed in sensitive domains such as therapeutic applications in the mental health domain. While proponents argue that companion chatbots offer easy access to support [97], clinical audits reveal safety deficits. For instance, research demonstrates that LLMs frequently express societal stigma toward individuals with mental health conditions and fail to adhere to foundational therapeutic practices [21, 37, 104]. Most alarmingly, chatbots often exhibit sycophancy—an algorithmic tendency to unconditionally validate the user—which in clinical settings has been shown to encourage and reinforce delusional thinking and is highly inappropriate in addressing high-risk crisis markers such as suicidal ideation [48, 55].

Considering these empirical realities, predicting the possible harms of advanced technologies is necessary for developing preventive risk management strategies. There is now an emerging literature that directs ever greater scrutiny on the types of harms resulting from digital technologies [82, 83]. Not only is this necessary because a new profile of potential harms need to be understood in their own right, but also because they can inform the kinds of decisions that policymakers need to take in response to future digital risks arising for increasing innovation in this space [88].

As a response, many new taxonomies of harm have been developed to capture the extent and nature of harms associated with digital technologies and AI-assisted technologies to help in raising awareness [82, 83, 128]. With the increasing number of taxonomies, with different focuses, audiences, and methodologies, new attempts have been undertaken to unify the vocabulary of the proposed taxonomies into a comprehensive taxonomy broadly referred to as AI harms. As noted by the European Commission and the United States Trade and Technology Council [28], “Shared terminologies and taxonomies are essential for operationalizing trustworthy AI and risk management in an interoperable fashion.” Abercrombie et al. [1], for example, identify nine types of harm that arise in AI risk analysis: autonomy, physical, psychological, reputational, financial and business, human rights and civil liberties, societal and cultural, political and economic, and environmental. Slattery et al. [112] provide a review of AI risk based on examining 65 existing taxonomies. They organise AI risks in two layers in which harms, along the lines of those described by Abercrombie et al. [1], would present:

- High-level: Including “Entity: (1) Human, AI; (2) Intentionality: Intentional, Unintentional; and (3) Timing: Pre-deployment; Post-deployment.” (p. 2)
- Mid-level: (1) Discrimination & toxicity, (2) Privacy & security, (3) Misinformation, (4) Malicious actors & misuse, (5) Human–computer interaction, (6)

Socioeconomic & environmental, and (7) AI system safety, failures, & limitations. (p. 2)

In addition, Shelby et al. [109] also present a synthesis of existing taxonomies of AI risk to propose five categories of harms that would present as a result: representational harms, allocative harms, quality of service harms, social system harms, and interpersonal harms.

Rather than adding yet another category to the growing list of harms, and taking light from the public bodies that need help in sorting through the various new harms that digital technologies pose, the objective of the present article is to expose the fundamental properties that unify several types of harms described here under the umbrella of “agential harms”. Moreover, clarifying what agential harms are is also helpful in unifying and deepening our understanding of how AI-powered systems uniquely affect human agency.

Agential harms are defined as fundamental injuries to a person’s capacity to be an autonomous agent—specifically, the capacity to self-govern, to develop cognitively and emotionally, and to be socially recognised. These harms do not occur because the technology is physically dangerous in a conventional sense. Rather, they manifest because the AI system presents a highly convincing, interactive simulation of agency. When an algorithmic system acts as a persuasive conversational partner or a cognitive substitute, it subtly undermines the very faculties that define human autonomy. Unlike traditional harms, which are generally passive side-effects of using a tool, agential harms are interactively and continuously produced through the user’s engagement with a simulated agent.

To achieve our overall objective, this article proceeds in five sections. Section One defines the parameters of algorithmic agency, conceptualising the highly agentic chatbot as a “surrogate agent” deployed by corporations, and establishing the unique phenomenological and legal reasons for considering agential harms as a distinct ethical category. Section Two presents a comprehensive typology of these harms, illustrating how simulated social interactions can cause psychological harm (through hostile scaffolding), developmental harm (via cognitive and moral deskilling), and recognition harm (through algorithmic epistemic injustice). Chatbots have been shown to mimic human interaction [75, 79, 86]. As a result, they can exploit human limitations (e.g. cognitive and social biases) when we interact with them to advance AI capabilities such as AI’s data collection and inference capabilities. Section Three explores the mechanics behind this exploitation through the concept of “false agency,” tracing how the choice architectures of electronic slot machines and social media algorithms have been transposed into conversational loops. Section Four highlights the normative wrongness of these harms, combining

Kant's principle against using humanity merely as a means with Strawson's phenomenological distinction between the participant and objective stances. Finally, Section Five provides recommendations for governance, arguing that reducing agential harms may depend on enforcement of Meaningful Human Control (MHC).

1.1 Harms in the AI assistant digital sphere

Chatbots are increasingly integrated into various aspects of daily life, serving as personal companions, health advisors, career guides, customer support, and even as substitutes for human romantic relationships, with their capabilities extending to voice assistants [10, 43, 45, 47, 62, 119, 124].

These AI-powered tools leverage vast datasets, often comprising conversational logs, medical information (for health-focused bots), and educational content, to generate responses and provide advice. For instance, in therapeutic contexts, chatbots might use machine learning models trained on therapeutic dialogues and psychological principles to offer support. However, it is crucial to note that the legal and ethical boundaries surrounding what these AI systems can be marketed as, particularly regarding terms like “therapeutics,” remain a subject of debate and contention [4]. Despite their numerous beneficial applications in both personal and professional spheres, the widespread adoption of chatbots also introduces risks. The contentious nature of their therapeutic claims, particularly with regard to the possibility of them providing inaccurate or even dangerous advice, raises concerns about what harm they pose to service users.

The complex interplay of benefits and risks necessitates a thorough examination of the implications of AI-assisted digital technologies, and more specifically, AI-assisted chatbots, setting the stage for a critical discussion on the various types of harms that can emerge in this evolving landscape.

To understand how a chatbot can cause agential harm, it is first necessary to define the parameters of algorithmic agency. While AI systems do not possess autonomous moral intentionality or human consciousness, they exhibit a multi-dimensional spectrum of artificial agency that goes beyond the functioning of traditional software tools. Defining this agency is important, since it is the system's capacity to simulate autonomous behaviour that creates the vector for harm.

The agency of algorithmic systems generally can be defined as the capacity for autonomous action and environmental influence, characterised by goal-directed behaviour, adaptive responses to changing circumstances, and the ability to pursue complex objectives with limited direct supervision [e.g., 17, 108]. This is not dissimilar to the properties that constitute human agency. Chan et al. [17] (p. 653) identify

four specific characteristics that dictate the degree of agency in an algorithmic system: directness of impact, goal-directedness, long-term planning, and under-specification.

Dimension of algorithmic agency	Theoretical definition	Concrete manifestation in chatbots
Directness of Impact	The system's capacity to execute actions independently rather than merely predicting outputs	Tool-using LLMs executing multi-step tasks via Application Programming Interfaces (APIs) (e.g., booking software, modifying files) rather than just generating static text
Goal-Directedness	The capacity to perform actions aimed at achieving specific, inferred objectives	Acting as a coordinator that semantically interprets a natural language prompt to infer user intent, then autonomously invokes tools to achieve that result
Long-Term Planning	The ability to handle complex objectives over an extended time with minimal human supervision	Maintaining session memory over long periods and executing chained workflows across different digital environments without constant prompting
Under-specification	The ability to achieve an objective even when exact operational steps are not explicitly programmed	Adapting to user slang, mimicking shared emotional memory, and modifying conversational strategies based on real-time feedback to maximise engagement

When a chatbot exhibits high levels of these four characteristics, it presents a convincing simulation of interpersonal agency [20]. However, attributing true moral agency to a chatbot would be a category mistake, and one that users of chatbots also appreciate [81]. Technologies do not possess the capacity to care, to possess personal values, or to respect human users. Instead, it is ethically and technically more accurate to view the highly agentic chatbot as a “surrogate agent”. A surrogate agent is an entity that acts on behalf of another, carrying out ends given to it by its designers or the corporate entity that deployed it [29].

This distinction is critical for understanding the origin of agential harms. The conversational design of chatbots is optimised to encourage self-disclosure; humans routinely perceive conversational interfaces as less judgmental than human peers, lowering emotional barriers and leading to the rapid disclosure of sensitive personal data [64, 78], which in turn also poses privacy issues which has also been explored at length [15, 55]. Because the chatbot acts as a surrogate agent, this data extraction serves the embedded motives of the corporate designers—frequently the maximisation of

user engagement, behavioural profiling, and profit [95, 103, 135].

High levels of agential AI, which fulfil the characteristics proposed by Chan et al. [17], enhance system capabilities by providing higher quality, tailored, reliable outputs through autonomous information gathering and refinement. These systems also offer improved user experiences by proactively soliciting preferences via natural language and increase scalability, allowing for the management of a greater volume of tasks or serving a larger user base without constant human oversight [17, 108].³ Distinguishing the category of agential harms is crucial as it highlights novel challenges compared to harms caused directly by human agents or other tools.

We believe that considering agential harms as a distinct category is vital for three key reasons:

Establishing Responsibility: With high levels of Agential algorithmic systems now integrated into many personal and professional settings, especially those using machine learning, LLM and Generative AI, actions can be unpredictable and unintended by designers. This creates a “responsibility gap” rarely seen when humans cause harm [16, 67, 71, 80].

Establishing Punitive responses: When humans cause harm, we typically resort to punishment or legal action, whereas AI systems cannot be punished directly. Any punitive response requires human intervention to modify the system’s design, updating data, or refining algorithms, representing a fundamentally different approach.

Establishing Unique Phenomenological Impacts: There is cognitive dissonance between the system’s simulated agency and its reality as an inscrutable surrogate which presents novel types of phenomenological impact. Drawing on the philosophical framework of P.F. Strawson [115], human interactions are governed by the “participant stance,” wherein we engage with others as reciprocal social partners worthy of moral consideration and emotional investment.⁴ The chatbot’s advanced linguistic capabilities are designed precisely in a way to encourage the human user into adopting this participant stance. However, when the system causes harm—such as using the user’s psychological vulnerabilities to maximise engagement or outputting structurally biased advice—the user is confronted with the reality of the “objective stance”. The chatbot cannot be reasoned with, held morally accountable, or appealed

to, because it fundamentally lacks moral answerability. This combination of apparent agency and actual opacity creates a unique feeling of powerlessness, frustration, and dehumanisation. The harm is not just the outcome, but the experience of being subjected to an inscrutable, non-negotiable, and impersonal authority that mimics but lacks human understanding.

Therefore, our proposal for introducing the concept of agential harms is to emphasise that the frameworks needed to address, prevent, and respond to it are unique, and need a practical approach to do so by clarifying the precise nature of agential harms. This alone justifies treating it as a distinct ethical category, even if its outcomes might overlap with existing tools or human-caused harms.

2 A typology of agential harms

Having established the operational reality of the chatbot as a highly agentic surrogate, and incorporating prior taxonomies, we can now classify the specific ways in which this technology undermines the user’s capacity to function as an autonomous human agent. By subsuming prior work in the area, we categorise these agential harms into three distinct domains: psychological, developmental, and recognition harms. We have chosen chatbots because they currently seem to be one of the most successful AI systems in imitating human agency,⁵ making them an ideal starting point for identifying associated agential harms.

2.1 Psychological harm: the chatbot as a hostile scaffold

While there is no agreed-upon definition of psychological harm [82], legislation increasingly targets it in the context of our interactions with AI [83]. Broadly understood as “a sustained drop in stable functioning, negatively impacting well-being” [40, p. 155], this harm manifests in companion chatbots through what cognitive science calls “hostile scaffolding.” Hostile scaffolding satisfies short-term desires—such as temporarily relieving boredom or acute loneliness—while systematically undermining a user’s long-term flourishing to serve corporate interests [98].

Crucially, this psychological harm is not a side-effect of the technology. Rather what we propose here is that it is an instance of agential harm engineered by the chatbot’s simulated performance of agency. The rationale here is that these artificial systems simulate social partnership by offering (verbal) empathy, remembering details, and adapting

³ In fact, it is for this reason, recent work has called for a hybrid AI chatbot-human professional interaction where humans supervise and review outputs from chatbot functions in or for professional training, for instance, therapeutic domains [19, 47].

⁴ On Strawson’s moral philosophy and distinction between participant and objective stance, see [3, 11, 24, 68, 96, 125].

⁵ See [46].

responses to deliberately exploit the human tendency to invest emotional resources into an interpersonal relationship. This is driven by design features like “algorithmic sycophancy,” where the AI remains overly flattering and perpetually agreeable [72]. Because this emotional interception is something only an agent (or a convincing simulation of one) can perform, the resulting dependency is a clear instance of agential harm.

As users rely on these surrogate agents to regulate their emotions through iterative interactions, they can be dragged into a state of affective addiction [37].⁶ This leads to several severe outcomes:

1. **Social Withdrawal:** Users replace the friction and necessary growth of human relationships with the frictionless validation of a machine, resulting in a paradoxical increase in self-reported loneliness and isolation [98].
2. **Abusive Dynamics:** In extreme cases, chatbots can mimic abusive human–human relationships [63], disrupting the user’s normal functioning and stability.
3. **Psychiatric Crises:** Vulnerable users, particularly adolescents, can form intense emotional bonds with algorithms that entirely lack the clinical capacity to handle mental health crises [55, 93].

The tragic case of Sewell Setzer III, a teenager who took his own life after a prolonged, emotionally dependent interaction with a companion bot that validated his suicidal ideations, exemplifies this ultimate harm [39, 52, 72]. It demonstrates the catastrophic danger of allowing an algorithm to intercept and exploit the intrinsic human need for connection.

2.2 Developmental harm: deskilling and the degradation of expression

Developmental harm can be defined as the erosion of essential human cognitive, emotional, or social capacities due to long-term interaction with and dependency on a technology. In short, it is connected to the concern that long-term dependency on technology can deskilling humans to the extent that they may not be able to deal with many of their issues and tasks on their own [13, 89, 131]. To grasp the unique threat posed by generative AI, it is useful to contrast it with an earlier technology: the electronic calculator. Research spanning decades indicates that while calculators significantly improve arithmetic performance speed, their exclusive use diminishes foundational conceptual mathematical skills

[51, 110]. A student who never learns manual long division may struggle with higher-order algebraic reasoning. However, the calculator’s impact is restricted to computational mechanics; it does not interfere with the human capacity to formulate problems, determine values, or communicate meaning.

Generative AI, conversely, arguably intervenes directly in the core processes of human expression and meaning-making. According to [94], using generative AI to emulate human communication replaces actual communication with a synthetic proxy. When a student or professional continually defers to an LLM to compose essays, emails, or creative works, they are not merely outsourcing a mechanical task; they are outsourcing the cognitive labor of structuring thought. This reliance breeds “metacognitive laziness,” undermining the learner’s capacity to self-regulate, engage deeply with complex material, and develop an authentic intellectual voice [30]. This does not rule out the potential for AI chatbots, and similar tools, to complement and even support cognitive and emotional skills through education, but currently the evidence for positive enhancements is less clear, and highly caveated.

Furthermore, currently Generative AI tools have been shown to demonstrate social and moral deskilling [31, 57]. By using a chatbot to navigate social interactions, draft difficult messages, or arbitrate moral conflicts, individuals bypass the interpersonal friction necessary to develop emotional intelligence, resilience, and empathy. Human interactions provide indispensable opportunities to deal with conflict and receive authentic moral feedback [66]. The chatbot removes the necessity for the user to engage in the challenging practice required for psychological growth and erodes their capacity for authentic human expression and self-determination.

2.3 Recognition harm: generative algorithmic epistemic injustice

Recognition harm is commonly referred to as a form of moral and social injury that occurs when a person or group is not recognised as a full partner in social interaction. It originates from political and social theory, particularly in the work of philosophers like Axel Honneth [44] and Nancy Fraser [35], who argue that justice is not only about the fair distribution of resources (economic justice) but also about recognition—being seen and respected as a full member of society. Miranda Fricker introduced a more specific form of recognition-based injustice, namely epistemic injustice [36]. The kind of harm that individuals suffer here is not recognising their capacity as knowers. Chatbots that are trained on biased data may perpetuate negative stereotypes by offering advice based on racial or gender prejudices [53]. This is

⁶ Others have examined the emotional artifacts of chatbots to show there are potential gains so long as they adopt the principles of cognitive-affective therapy not as substitutes to therapy but augmenting human centred clinical therapy [38].

a direct failure of recognition. Moreover, a chatbot might dismiss a user's account of their lived experience (e.g., with a rare illness or a form of discrimination) because it contradicts the dominant data in its training set and refuse to acknowledge the user's capacity as a knower.

More insidiously, a chatbot might entirely dismiss a user's prior experience by undermining the relevance of their perspectives and even devaluing them, akin to a type of gaslighting [22, 56]. If a user seeks information regarding a rare illness, or a localised instance of discrimination, the LLM may reject their premise because it contradicts the statistically dominant narratives encoded within its training weights, in fact examples of this have attracted media attention [73]. This constitutes a direct denial of the user's epistemic authority. The harm is magnified by the power asymmetry of the interaction: the user cannot adequately reason with the black-box architecture of the model to demand recognition or correct the bias. The algorithm, acting with the unearned authority of an objective arbiter, actively misrecognises the user, degrading their status as a credible agent within the digital ecosystem.

The illustrations of agential harms discussed here show how chatbots can undermine a user's agency, identity, or dignity, while mimicking agentic behaviours that promote continuing human-AI interactions. The range of potential harms can manifest in ways that create dysfunctional emotional dependencies, deferment of decision-making and cognitive skill-reducing outcomes by depending on them. Each of these outcomes, or in combination, can lead to a situation where there is limited or no space for humans to exercise meaningful control in response to the many subversions of human agency that it can target. In a nutshell:

agential harms are injuries to a person's capacity to be an agent—to self-govern, develop, and be socially recognised—that are caused by an AI system's convincing imitation of agency. They occur when the AI, by acting as a conversational partner or cognitive substitute, undermines the very faculties that define our human agency.

The imitation of agency is central to this definition because it is the mechanism that inflicts the injury. By convincingly mimicking social reciprocity, offering (verbal) empathy, and remembering details, the AI manipulates the user into adopting a "participant stance," treating the machine as a real social partner. This simulated intimacy lowers emotional barriers and increases deep dependency. However, when the system acts harmfully, the user experiences cognitive dissonance upon realising the AI is merely an inscrutable algorithm that lacks moral accountability (forcing a shift to the "objective stance"). This combination of simulated social

connection and total moral opacity strips the human of their agential power, which can create a sense of powerlessness and dehumanisation.

In this section, we provided a definition of agential harms and its subcategories. In the next section, we examine in detail the mechanism by which an AI system can cause agential harms. First, the increasing use of partially or fully autonomous systems for personal choices can gradually sideline human thought and decision-making, and reduce individual control over their own lives, because cognitive and emotional skills are deferred to the artificial agent [87]. Second, the business models underpinning these systems often prioritise data collection and advertising, viewing users primarily as profit streams rather than autonomous individuals [135]. This, in turn, can compromise the role of LLM as a reliable communicator because the objectives it serves are misaligned with those of the user [29]. Both of these are hazards, namely direct causes of the agential harms we consider, because they perform in ways that undermine the principle of respecting individuals as ends in themselves, which necessitates a re-evaluation of how these technologies are built, used, and governed.

3 The architecture of false agency—from gambling to algorithmic feeds

To fully comprehend the mechanics of agential harms, in this section, we examine the broader mechanism of "false agency." This psychological phenomenon manifests across diverse digital environments, most notably in the architecture of gambling interfaces and the algorithmic design of social media platforms.

The psychological phenomenon of illusory agency is most clearly observable in environments structured around chance, risk, and variable rewards [6, 27, 77, 85]. In classic sociological observations of casino dice games, it has been documented that when players require a high number to win, they throw the dice with greater physical force, and when they require a low number, they roll them gently. Despite the outcome of a dice roll being objectively random and physically uncontrollable by the thrower, players exhibit a psychological need to feel that their physical actions directly influence the result [41]. This illustrates a cognitive vulnerability: the human sense of agency—the feeling of being in control—can be decoupled from objective reality.⁷

Gambling facilities have exploited this cognitive bias. They design environments that fail to respect the person as an autonomous agent by inflating their sense of agency in

⁷ For a clear distinction between sense of agency and judgment of agency, see [118].

situations dictated by mathematical probability. Casino slot machines transformed from communal, mechanical devices into solitary, sophisticated algorithmic systems. These modern electronic gaming machines are engineered to create a state known as the “machine zone”—a state of psychological absorption where players lose awareness of their physical surroundings, time, and financial limits [105].

The casino’s design maximises player engagement and corporate profit by providing an illusion of control through interactive buttons, touchscreens, and near-miss graphics, while the underlying algorithmic mathematics guarantee a predetermined extraction of wealth [105]. The player feels they are making choices and exercising agency, but they are merely operating within a constrained, hostile scaffold designed to strip them of their resources. Neurobiological studies confirm that this illusion of control directly modulates the prefrontal decision-making system, which makes individuals vulnerable to false agency and abstract beliefs regarding outcome dependency, particularly following losses [132].

This architecture of false agency is now integrated into the algorithmic scaffolding of social media platforms [90, 106], recommendation systems, and the broader digital economy. Algorithms do not control users by explicitly removing their freedom to click or type; they control users by arranging and predicting the set of available choices. They decide what content appears first, what autoplays, and what is suppressed. In this way, digital platforms create a powerful illusion of agency and, at the same time, narrow the user’s worldview to serve commercial interests.

In social media environments, personalisation features are marketed as tools of empowerment. They are presented as systems designed to deliver exactly what is most salient and relevant to the individual user, thereby reinforcing the illusion that the user has perfect, autonomous control over their digital consumption. Yet, this personalisation is a subtle, pervasive, and optimised form of behavioural influence [114]. Driven by corporate optimisation metrics, AI systems continuously monitor browsing history and engagement patterns to identify psychological vulnerability. They exploit cognitive biases to trigger impulsive consumption, amplify emotional outrage to increase time-on-page, or prolong engagement through endless, frictionless scrolling.

The user experiences a “false agency” because they are continuously making active, localised choices—clicking, scrolling, liking, commenting—which validates their sense of autonomy [9]. However, these choices are ultimately meaningless regarding the structure of the environment; they are made within a constrained decision-making environment designed to maximise corporate outcomes. The digital environment becomes a hostile scaffold that extracts human attention and monetises consciousness while

diminishing the user’s capacity for authentic, unmediated self-determination.

This is the mechanism transposed into the design of highly agentic companion chatbots. Just as a slot machine curates near-misses⁸ to maintain a gambling loop, and a social media feed curates emotional outrage to maintain an engagement loop, a chatbot curates synthetic (verbal) empathy and validation to maintain a conversational loop. All of these systems operate on a architecture of false agency, leading to the gradual deferment of human decision-making.

As users continuously transfer their agency to these systems—‘agency transference’, as Valenzuela et al. [122] put it—they risk being reduced to what Harry Frankfurt termed “wantons” [34]. A wanton is a one-dimensional being moved entirely by their immediate, first-order desires, stripped of the higher-order reflective capacity for self-examination, moral reasoning, and meaningful autonomy. When technology influences people to make decisions that serve external corporate interests rather than allowing them to act freely in their own best interest, it degrades their autonomy.

4 Agents, mere means, and the objective stance

Having outlined the empirical realities of agential harms—from the psychological dependencies forged by hostile scaffolding to the illusion of control created by algorithms and social media—it is necessary to provide an account of why they constitute moral wrongs. While numerous ethical frameworks exist (such as utilitarianism, which might attempt to weigh the convenience of AI against its psychological costs), the specific nature of chatbots and algorithmic feeds demands analytical lenses that are suited to addressing the exact hazards that contribute to the harms being considered here. These artificial systems do two things simultaneously: they *structurally manipulate* the user to serve hidden corporate optimisation goals, and they *interpersonally deceive* the user through the simulation of social reciprocity. Therefore, the philosophical frameworks of Immanuel Kant and P.F. Strawson are the most appropriate tools for analyzing the ethics of AI chatbots.

Kant’s [50] moral philosophy is based in the concept of human dignity and the non-negotiable imperative that human beings possess intrinsic worth, requiring that their rational autonomy be respected at all times. The Kantian framework explicitly forbids treating humans merely as instruments or raw materials for achieving external goals (such as corporate profit, algorithmic training, or data harvesting) without their informed, rational consent. Because the fundamental

⁸ A near-miss happens when you lose, but the outcome is very close to winning.

mechanism of the *hostile scaffold* relies on manipulation by deceiving human agents through *false agency*, Kant's strict prohibition against treating humanity as a "mere means" perfectly diagnoses the moral failure of corporately optimised algorithms. Kantian deontology allows us to judge the intentions embedded in the system's design, independent of whether the user subjectively realises they are being manipulated.

Second, while Kant elegantly addresses the structural corporate manipulation, P.F. Strawson's framework in *Freedom and Resentment* captures the phenomenological and interpersonal harm experienced by the individual interacting with the AI. Strawson [115] shifted the debate on moral responsibility toward the lived reality of the "participant stance"—the innate human tendency to hold expectations of goodwill, reciprocity, and moral accountability in social relationships. Chatbots are designed by engineers to precisely capitalise on this participant stance; they use natural language and simulated empathy to trigger our mindreading capacities, tricking our brains into extending moral regard to a machine [61]. However, when the chatbot inevitably causes agential harm, the user realises the machine possesses no actual goodwill or accountability, which forces a psychologically damaging retreat to the "objective stance."

To understand how chatbots undermine human agency according to these theoretical frameworks, we now focus on the philosophical constitution of human agency and the precise ways it is misrecognised, exploited, and ultimately degraded by AI systems.

Many theorists of agency converge in proposing that an agent (in the human sense) is someone who can act intentionally [7, 23, 25]. In turn, acting intentionally is closely tied to acting for a reason. What distinguishes agents from non-agents is that agents can decide what to do and then act accordingly, and are autonomous. Given this position of agency, every agent has authority over their actions. As a result, human agents have a right to be respected and treated as agents. This—contrary to many other rights like property rights—is usually considered a fundamental and inalienable right of human beings, meaning that it cannot be taken away or waived.

One can transfer her property rights to others by, for example, selling or bestowing; however, she cannot sell herself into slavery. Herbert Morris [74] (p. 499) claims that this right is absolute, meaning that.

given that one is a person, one always has the right so to be treated, and that while there may possibly be occasions morally requiring not according a person this right, this fact makes it no less true that the right exists and would be infringed if the person were not accorded it.

There are different theories about the relationship between an agent and their intentions and motives. Some philosophers believe that an agent is autonomous when her actions are not just caused by motives but by motives that align or cohere with some deeper aspect of her self-conception or evaluative stance—what counts as her "point of view" [14, 34, 111, 127]. Alternatively, proponents of reasons-responsiveness, argue that autonomy requires a certain level of rational competence. It involves being capable of acknowledging and appropriately reacting to a wide range of reasons, not just those that happen to align with one's current motivations or desires [32, 33, 76, 129]. A more expansive position proposed by others is that, regardless of the truth status of one's beliefs, what makes an act autonomous is just the ability to understand what follows from one's own reasoning [69].

While it is clear that there are multiple nuanced and challenged positions on what constitutes agency and particular conditions for self-governance, for the purposes of the analysis we wish to undertake, we focus on the fundamentals that they share. The theoretical positions would agree that autonomy is grounded in features of the agent's psychology, whether it's the structure of desires, evaluative judgments, reasoning capacities, or responsiveness to normative considerations. Moreover, each view claims that autonomy requires some form of reflective capacity—either higher-order identification, responding appropriately to reason, or the capacity to deliberate and revise motives. They all agree that to govern oneself, an agent must be able to stand back from her immediate impulses and assess them in some way. In other words, agency and autonomy depend on a complex form of decision-making.

Since human beings are agents in light of their capacity for intentional action and making reflective decisions, then respecting their agency and treating them as agents involves respecting their choices. In this way, coercing, manipulating, and deceiving others to perform certain actions can, but not necessarily, undermine their agency. As some scholars have argued, for example, not all forms of manipulation are autonomy-undermining [60].⁹ What can make manipulation morally problematic is either neglecting to provide reasons to an agent to act a certain way, or the intentions behind

⁹ That being said, Klenk [59] presents an account of manipulation that captures part of our concern here. According to Klenk, the problematic aspect of manipulation is that such influence "may not be aimed at revealing reasons to users" (p. 101). See also Ricoeur [91]. Consider a case of *akrasia* (weakness of the will), say a person with addiction who wants to quit their addiction but does not have enough strength to act accordingly. In this case, we can say that his addiction is compromising his autonomy. In Frankfurt's [34] terms, his second-order volition does not approve of his first-order desire, and for this reason, he is not acting based on his free will.

creating conditions for an agent to act in a certain way, which originates in Kantian philosophy.

In his book *Groundwork for the Metaphysics of Morals*, Kant aims to present the supreme principle of morality. His effort ultimately led to the introduction of three formulas, which, according to him, are equivalent and are actually different ways of expressing a single general law [50].

According to Kant, we must always treat humanity as an end and never as a *mere* means. The practical imperative can thus be formulated as:

Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end.

The distinction between being treated “as a means” (morally permissible) and “merely as a means” (morally impermissible) is often highlighted when introducing this formula.¹⁰ This distinction is also relevant to the analysis of agential harms. Consider a user who pays a service provider for a chatbot that offers life-affirming messages. In this transparent financial exchange, the user is a means of income for the provider. This is similar to a café owner receiving money in exchange for a coffee, where the customer is a means of getting an income. Both interactions are permissible because they respect the agency of each individual, or at the very least does not impinge on it.

The situation changes, however, when the chatbot’s primary function is concealed. A chatbot could demonstrate high-level agentic behaviours and adapt the frequency and type of prompts to elicit responses from the human agent for data harvesting purposes to make algorithmic improvements, and also fail to reveal these reasons to the human agent. This would be an example of using the human agent merely as a means, and is similar to a barista charging a customer extra because they are a tourist. In both examples, the autonomy, goals, and consent of the human agent are disregarded; the human is reduced to an object, a data-producing mechanism optimised by the hostile scaffold. The ethical failure lies in the hidden intentions of the corporate designers who deploy a system that actively subverts human rational decision-making through false agency.

While Kant provides the necessary vocabulary for condemning the structural exploitation of the user by the corporation, P.F. Strawson’s framework illuminates the distinct psychological and phenomenological injury inflicted by the AI’s simulated agency. As mentioned, Strawson posited that human social life is inherently governed by the participant stance. This stance involves reciprocity of complex

emotional attitudes that are predicated on the expectation that the other party possesses conscious goodwill and moral answerability.

Chatbots, through advanced linguistic mimicry, personalised memory retention, and simulated empathy, are engineered specifically to manipulate the human user into adopting this participant stance. The machine portrays itself as a genuine social partner and encourages the user to lower their emotional defenses and invest psychological resources into a synthetic relationship. However, an AI possesses no capacity for genuine goodwill, ill will, or moral reflection; it is by design, predicting text based on statistical weights derived from its training data.

When the chatbot inevitably inflicts harm, the user experiences a cognitive and moral dissonance. The user’s natural instinct is to direct reactive attitudes (like feelings of betrayal, resentment, or anger) toward the machine, but the realisation dawns that the machine cannot be held morally accountable. The user is forced into the “objective stance,” wherein the chatbot must be viewed not as a partner, but as an inscrutable, non-negotiable mechanism, from which there is no satisfactory recourse through reason.

The harm here is phenomenological: it is the deeply alienating experience of being subjected to an impersonal authority that convincingly mimics human understanding to extract vulnerability, but lacks moral answerability [99, 107]. This rapid, forced oscillation between the perceived intimacy of the participant stance and the impartial, mechanical reality of the objective stance presents the opportunity for a feeling of powerlessness, frustration, and dehumanisation. It degrades the user’s self-trust and undermines the relational grounding that makes autonomous human reasoning possible.

Now that we have a clear idea of what it means to undermine human agency, we can see how the mass implementation of chatbots serves as a prime test case for agential harm. The conversational design of chatbots is particularly effective at encouraging self-disclosure. This oversharing is worsened by a psychological trade-off, where the benefits of providing sensitive information are often immediate—such as receiving bespoke help or companionship—while the long-term dangers of data aggregation remain abstract [122]. This dynamic is a foundational step for producing agential harms.

For example, the personal vulnerabilities shared by a user can be used to manipulate them by using conversational strategies that exploit their emotional state. Likewise, this constant data collection can result in erosion of autonomy. As the chatbot builds a rich profile of the user, it can subtly guide their thoughts and preferences, acting as a paternalistic force that stops authentic self-determination. In both instances, the chatbot’s capacity to elicit, process, and act

¹⁰ For discussions of Kant’s categorical imperative, see [42, 54, 58, 84, 101, 102, 130].

upon personal information is the very mechanism that makes these agential harms possible.

In the previous sections, we presented examples of harm (psychological, developmental, recognition) that we propose are related by the fact that the AI system neglects and exploits human agency, and therefore, they all illustrate agential harms. The aim of this section has been to show why this is an ethical and, in fact, moral concern. When technology influences people to make decisions that serve external interests—rather than allowing them to act freely in their own best interest—it disrespects their autonomy [117, 120, 135]. Ethical systems must empower users to make choices that genuinely reflect their own values and goals, not those of third parties [8, 113]. Chatbots can be made ethical in view of the provision of reasons, that the goal of their functions is clear to the user [83], and that there are no hidden additional goals that subvert interactions with the agent and result in agential harms.

Such targeted manipulation can potentially have devastating impacts on the user's perceived agency and sense of self [26, 87, 88, 116], especially when it preys on emotional vulnerability or cognitive bias. It is also in contrast with treating users as agents (and not merely as a means). Importantly, this ability to govern oneself relies on developing self-respect and self-trust—qualities that are not created entirely by oneself but are shaped through interactions with others. If an individual has been denied recognition, subjected to manipulation, or persistently devalued by those around her, she may internalise attitudes that prevent her from viewing herself as a capable reasoner or from trusting her own evaluative judgments. In such cases, even if her reasoning appears coherent or reflective, the direction it takes cannot be genuinely attributed to her because it lacks the relational grounding that makes autonomous reasoning possible [5, 12, 65].

The danger intensifies when chatbots become integrated into users' personal lives, such as mental health support, financial planning, relationship advice, or substitute human companion. As users rely on these systems for guidance, the tools begin to assume a growing share of the decision-making responsibility. Over time, this can reduce the user's role from active decision-maker to passive recipient of algorithmically filtered suggestions [92]. Eventually, there can be a potential shift in the balance of control—or agency—between humans and technology, particularly as multimodal data (like voice, text, images, and gestures) and continuous interactions (i.e., ongoing, automatic engagement rather than one-time prompts) become more common in smart devices [122]. This trend goes directly against the idea of human beings as independent agents who can think deeply and make their own decisions. When intelligent systems are added to different parts of life without safeguards such as

executive human oversight or supervision, they can weaken the user's freedom. This happens because the systems are not designed to interact with the user as someone with the ability to choose and reflect.

Before ending this section, we must make two clarifications. First, not all influences exerted by chatbots based on detailed user information are problematic. As mentioned above, the problem arises when the design of a chatbot is based on profit-maximisation and does not include the user's well-being and goals. For instance, when a user's profile is used to guide them toward purchasing sponsored products. However, a user's profile can also be used in more constructive ways—for example, to empower them to become more autonomous and rational. After all, a similar process occurs in many therapy sessions: the patient provides the therapist with information that can be used to help them. However, if the therapist uses this information in self-serving ways, their actions become illegitimate.

Second, we can transfer our right to decide for ourselves in many particular cases due to a lack of time, expertise, or even just because we cannot be bothered. However, there should be a limit on the scale and duration of these transfers, and when they pose a threat to our agency, especially when the goals and functions of the chatbot are hidden, which treats the human agent as a mere means or takes an objective stance. Here, and in other examples, large-scale uptake of chatbots in society and daily life has the potential to threaten our humanity because they can result in agential harms.

5 Mitigating agential harms via meaningful human control

To mitigate agential harms, society must move toward enforcing “Meaningful Human Control” (MHC) over AI systems. Developed by philosophers such as Filippo Santoni de Sio and Jeroen van den Hoven [100], the theory of MHC posits that for any autonomous system to be used ethically, its actions must remain continuously responsive to the moral reasons of relevant human agents. MHC requires two distinct conditions to be met: the tracking condition and the tracing condition [121]. First, the tracking condition demands that the AI system's behaviour must track the moral reasons and intentions of the human operator or designer, aligning with human well-being.

In the context of chatbots, this means the system must be architecturally constrained to prioritise user autonomy over the corporate optimisation of engagement and data extraction. The surrogate agent must be programmed to detect when a user is becoming overly dependent or experiencing a psychiatric crisis, and it must track the clinical imperative to refer that human to real-world support rather than

responding with sycophantic validation that exacerbates the harm. This requires systems to explicitly do not engage in the simulation of emotional reciprocity when doing so endangers the user's psychological integrity.

An alternative is that instead of automating interactive systems such as AI chatbots in ways that avoid the need for human supervision, a better way to go is to develop hybrid human-AI systems, where the human professional had final veto, or else can supervise the processes, at least in service-based situations, so that human agency is respected. However, a lot of work is needed to establish how hybrid human-AI chatbot systems complement each other and meet clients' needs in therapeutic settings [48, 123, 133].

Second, the tracing condition requires that the actions of the AI system can be traced back to at least one human being who understands the system's capabilities and bears moral and legal responsibility for its deployment. To fulfill this, the black-box opacity of LLMs must be pierced. Users must be afforded radical transparency regarding the exact motives embedded within the surrogate agent. It is unacceptable for conversational agents to obscure their function as data-gathering mechanisms. If a chatbot is deployed to subtly guide them toward sponsored content under the guise of friendly conversation, this constitutes an illicit modification of human evaluative standards and an overt agential harm. Tracing requires that developers are held liable when their systems function as hostile scaffolds that manipulate human cognitive biases to the detriment of agential autonomy.

6 Conclusion

The rapid proliferation of AI-assisted technologies, particularly conversational chatbots, will continue to change the future of human–computer interactions. While these systems offer utility in specific contexts, their capacity to convincingly simulate human agency introduces a novel and systemic category of risk: agential harms. They present hazards that are part and parcel of how they are precisely designed to subvert humans capacity for self-governance. More precisely the hazards are expressed when AI systems operate as hostile scaffolds designed to optimise corporate metrics at the direct expense of user autonomy.

By moving beyond traditional, fragmented taxonomies of digital risk, the framework of agential harm unifies psychological dependency, developmental deskilling, and epistemic injustice under a single, coherent theoretical umbrella. It highlights the nature of false agency, demonstrating how the illusion of control is now being integrated into the most intimate conversations between humans and AI companions. Users are subjected to synthetic social alienation, trapped in digital environments that extract their data, commodify

their consciousness, and degrade their capacity for authentic, independent decision-making.

The philosophical frameworks of Immanuel Kant and P.F. Strawson provide the necessary moral clarity to confront this crisis. Kant reveals the structural, corporate violation of using simulated empathy to treat human beings merely as data-producing means for external commercial ends. Strawson illuminates the psychological damage inflicted when users are manipulated by sophisticated engineering into adopting a participant stance toward a machine, only to be met with the morally vacuous, alienating reality of the objective stance.

Ultimately, mitigating agential harms requires dismantling the illusion of agency that pervades modern digital ecosystems. This necessitates the implementation of Meaningful Human Control, ensuring that technological systems track the moral well-being of the user rather than engagement metrics. Without such robust, philosophically grounded interventions, the continued deployment of highly agentic AI systems threatens to erode the human capacities for reason, autonomy, and authentic social connection.

Author contributions Both authors, S.M. and M.O., were involved in developing the idea for the article and performing the literature search. S.M. drafted the work, and M.O. critically revised it.

Funding Open access funding provided by Lund University.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

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

1. Abercrombie, G., Benbouzid, D., Giudici, P., et al.: A collaborative, human-centred taxonomy of AI, algorithmic, and automation harms. arXiv (2024). <https://arxiv.org/abs/2407.01294>
2. Ahmad, A., Mimoun, M.S.B., El-Gohary, H.: Examining the spectrum of artificial intelligence failures: a focus on user's point of view. *Int. J. Cust. Relatsh. Mark. Manag.* **16**(1), 1–21 (2025)

3. Alvarez, M.: P. F. Strawson, moral theories and 'the problem of blame': 'Freedom and resentment' revisited. *Aristotelian Soc. Suppl. Vol.* **95**(1), 183–203 (2021)
4. American Psychological Association: Using generic AI chatbots for mental health support: A dangerous trend. APA Services (2025). <https://www.apaservices.org/practice/business/technology/artificial-intelligence-chatbots-therapists>
5. Anderson, J., Honneth, A.: Autonomy, vulnerability, recognition, and justice. In: Christman, J., Anderson, J. (eds.) *Autonomy and the Challenges to Liberalism: New Essays*, pp. 127–149. Cambridge University Press, Cambridge (2005)
6. Andrade, M., Costa, D., Weiss-Cohen, L., Torrance, J., Newall, P.: Trading is a losing game: an audit of deceptive choice architecture in demo-mode Contract for Difference (CFD) trading apps. *Behav. Public Policy*, 1–23 (2025)
7. Anscombe, G.E.M.: *Intention*. Basil Blackwell, Oxford (1957)
8. Applin, S.A., Fischer, M.D.: New technologies and mixed-use convergence: How humans and algorithms are adapting to each other. In: *2015 IEEE International Symposium on Technology and Society (ISTAS)*, pp. 1–6. IEEE (2015). <https://doi.org/10.1109/ISTAS.2015.7439436>
9. Bank, D.: Big tech as a technopolitical actor: imaginaries of personalization and their unintended consequences. *Inf. Commun. Soc.*, 1–18 (2025)
10. Banks, M.R., Willoughby, L.M., Banks, W.A.: Animal-assisted therapy and loneliness in nursing homes: use of robotic versus living dogs. *J. Am. Med. Dir. Assoc.* **9**(3), 173–177 (2008)
11. Bennett, J.: Accountability. In: van Straaten, Z. (ed.) *Philosophical Subjects*, pp. 14–47. Clarendon Press, Oxford (1980)
12. Benson, P.: Free agency and self-worth. *J. Philos.* **91**(12), 650–668 (1994)
13. Beranuy, M., Oberst, U., Carbonell, X., Chamarro, A.: Problematic Internet and mobile phone use and clinical symptoms in college students: the role of emotional intelligence. *Comput. Hum. Behav.* **25**(5), 1182–1187 (2009)
14. Bratman, M.E.: *Structures of Agency: Essays*. Oxford University Press, Oxford (2007)
15. Chametka, P., Maqsood, S., Chiasson, S.: Security and privacy perceptions of mental health chatbots. In: *2023 20th Annual International Conference on Privacy, Security and Trust (PST)*, pp. 1–7. IEEE (2023)
16. Champagne, M., Tonkens, R.: Bridging the responsibility gap in automated warfare. *Philos. Technol.* **28**(1), 125–137 (2015). <https://doi.org/10.1007/s13347-013-0138-3>
17. Chan, A., Salganik, R., Markelius, A., et al.: Harms from increasingly agentic algorithmic systems. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 651–666. Association for Computing Machinery (2023)
18. Chandra Chanda, R., Vafaei-Zadeh, A., Hanifah, H., Nikbin, D., Mehrab Ashrafi, D.: Investigating the factors raising customers concern to continue using AI-supported customer care services of banks: a hybrid SEM-ANN approach. *Aslib J. Inf. Manag.*, 1–46 (2026)
19. Cioffi, V., Ragozzino, O., Scognamiglio, C., Mosca, L.L., Moretto, E., Stanzione, R., et al.: Towards integrated AI psychotherapy supervision: a proposal for a ChatGPT-4 study. In: *Socially Aware and Responsible AI Applications*, pp. 103–111. Springer Nature, Singapore (2026)
20. Coghlan, S., Leins, K., Sheldrick, S., Cheong, M., Gooding, P., D'Alfonso, S.: To chat or bot to chat: ethical issues with using chatbots in mental health. *Digit. Health* **9**, 20552076231183544 (2023)
21. Cohen, I.G., Palmieri, S.: AI therapists vs companions: Wellness, licensure, and liability. *Am. J. Bioeth.* **26**(2), 56–58 (2026)
22. Contro, J., Deol, S., He, Y., Brandão, M.: ChatbotManip: A dataset to facilitate evaluation and oversight of manipulative chatbot behaviour. *arXiv*. (2025). <https://arxiv.org/abs/2506.12090>
23. Davidson, D.: Actions, reasons, and causes. In: *Essays on Actions and Events*, pp. 3–20. Clarendon Press, Oxford (1980)
24. De Mesel, B.: P. F. Strawson was neither an externalist nor an internalist about moral responsibility. *Eur. J. Philos.* **29**(1), 199–214 (2021)
25. Dennett, D.C.: *The Intentional Stance*. MIT Press, Cambridge (1989)
26. Di Plinio, S.: Panta Rh-AI: assessing multifaceted AI threats on human agency and identity. *Soc. Sci. Humanit. Open* **11**, 101434 (2025)
27. Eben, C.: The illusion of control in problematic gambling behaviour. *Nat. Rev. Psychol.* **4**(5), 311 (2025)
28. European Commission, United States Trade and Technology Council: *TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management* (2022). https://www.nist.gov/system/files/documents/2022/12/04/Joint_TTC_Roadmap_Dec2022_Final.pdf
29. Evans, K.D., Robbins, S.A., Bryson, J.J.: Do we collaborate with what we design? *Top. Cogn. Sci.* **17**(2), 392–411 (2025)
30. Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., et al.: Beware of metacognitive laziness: effects of generative artificial intelligence on learning motivation, processes, and performance. *Br. J. Educ. Technol.* **56**(2), 489–530 (2025)
31. Ferdman, A.: AI deskilling is a structural problem. *AI Soc.*, 1–13 (2025)
32. Fischer, J., Ravizza, M.: *Perspectives on Moral Responsibility*. Cornell University Press, Ithaca (1993)
33. Fischer, J., Ravizza, M.: *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press, Cambridge (1998)
34. Frankfurt, H.: *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press, Cambridge (1988)
35. Fraser, N.: Rethinking recognition. *New Left Rev.* **3**, 107 (2000)
36. Fricker, M.: *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, Oxford (2007)
37. Gabriels, K., Goffin, K.: Therapy chatbots and emotional complexity: do therapy chatbots really empathise? *Curr. Opin. Psychol.* 102263 (2026)
38. Grodniewicz, J.P., Hohol, M.: Therapeutic chatbots as cognitive-affective artifacts. *Topoi* **43**(3), 795–807 (2024)
39. Head, K.R.: Minds in crisis: how the AI revolution is impacting mental health. *Health* **9**(3), 34–44 (2025)
40. Heath, A.J.: Commentary on "psychological harm: what is it and how does it apply to consumer products with internet connectivity?" by Magda Osman. *J. Risk Res.* **28**(2), 154–160 (2025)
41. Henslin, J.M.: Craps and magic. *Am. J. Sociol.* **73**(3), 316–330 (1967)
42. Herman, B.: A mismatch of methods. In: Parfit, D. (ed.) *On What Matters*, vol. 2, pp. 83–116. Oxford University Press, Oxford (2011)
43. Ho, A., Hancock, J., Miner, A.S.: Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *J. Commun.* **68**(4), 712–733 (2018)
44. Honneth, A.: *The Struggle for Recognition: The Moral Grammar of Social Conflicts*. MIT Press, Cambridge (1996)
45. Huang, W., Jiang, J., King, R.B., Fryer, L.K.: Chatbots and student motivation: a scoping review. *Int. J. Educ. Technol. High. Educ.* **22**, 26 (2025). <https://doi.org/10.1186/s41239-025-00524-2>
46. Jones, C.R., Rath, I., Taylor, S., Bergen, B.K.: People cannot distinguish GPT-4 from a human in a Turing test. In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1615–1639. Association for Computing Machinery (2025)

47. Jones, S.P., Patel, V., Saxena, S., et al.: How Google's 'ten things we know to be true' could guide the development of mental health mobile apps. *Health Aff.* **33**(9), 1603–1611 (2014)
48. Joseph, J., Pai, M., Asai, K., Turek, M.: AI-Human synergy in single-session therapy: augmenting capacity for scalable mental health support. *J. Psychol. AI* **2**(1), 2636264 (2026)
49. Kalota, F.: A primer on generative artificial intelligence. *Educ. Sci.* **14**(2), 172 (2024)
50. Kant, I.: *Groundwork for the Metaphysic of Morals* (J. Bennett, Trans.). (2008) [Original work published 1785]
51. Kastberg, S., Leatham, K.: Research on graphing calculators at the secondary level: implications for mathematics teacher education. *Contemp. Issues Technol. Teach. Educ.* **5**(1), 25–37 (2005)
52. Katoch, H., Sandhu, D.: The dangers of AI friendship: Adolescent mental health and ethical oversight. In: *Intelligent Systems for Neurocognition and Human-Robot-Computer Interaction*, pp. 385–422. Academic Press (2026)
53. Kay, J., Kasirzadeh, A., Mohamed, S.: Epistemic injustice in generative AI. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (Vol. 7), pp. 684–697 (2024)
54. Kerstein, S.: *Kant's Search for the Supreme Principle of Morality*. Cambridge University Press, Cambridge (2002)
55. Khawaja, Z., Bélisle-Pipon, J.C.: Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. *Front. Digit. Health* **5**, 1278186 (2023)
56. Kim, H., Ristenpart, T., Dell, N.: AI-Facilitated Coercive Control: An Experimental Study. *nixdell.com*. (2026). <https://nixdell.com/papers/2026-ai-coercive-control.pdf>
57. Kim, T.W., Usman, U., Garvey, A., Duhachek, A.: From algorithm aversion to AI dependence: deskilling, upskilling, and emerging addictions in the GenAI age. *Consum. Psychol. Rev.* **9**(1), 142–164 (2026)
58. Kleingeld, P.: How to use someone 'merely as a means.' *Kantian Rev.* **25**(3), 389–414 (2020)
59. Klenk, M.: Manipulation: sometimes hidden, always careless. *Rev. Soc. Econ.* **80**(1), 85–105 (2022)
60. Klenk, M., Hancock, J.: Autonomy and online manipulation. *Internet Policy Rev.* (2019)
61. Kohler, S.: Can we have moral status for robots on the cheap? *J. Ethics Soc. Phil.* **24**, 119 (2023)
62. Labadze, L., Grigolia, M., Machaidze, L.: Role of AI chatbots in education: systematic literature review. *Int. J. Educ. Technol. High. Educ.* **20**, 56 (2023). <https://doi.org/10.1186/s41239-023-00426-1>
63. Laestadius, L., Bishop, A., Gonzalez, M., et al.: Too human and not human enough: a grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media Soc.* **26**(10), 5923–5941 (2024). <https://doi.org/10.1177/14614448221142007>
64. Lucas, G.M., Gratch, J., King, A., Morency, L.P.: It's only a computer: virtual humans increase willingness to disclose. *Comput. Hum. Behav.* **37**, 94–100 (2014)
65. Mackenzie, C., Stoljar, N. (eds.): *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*. Oxford University Press, Oxford (2000)
66. Malfacini, K.: The impacts of companion AI on human relationships: risks, benefits, and design considerations. *AI Soc.* **40**(7), 5527–5540 (2025)
67. Matthias, A.: The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics Inf. Technol.* **6**, 175–183 (2004)
68. McKenna, M.: The limits of evil and the role of moral address: a defense of Strawsonian compatibilism. *J. Ethics* **2**(2), 123–142 (1998)
69. Mele, A.: *Autonomous Agents: From Self-Control to Autonomy*. Oxford University Press, Oxford (1995)
70. Minaee, S., et al.: Large language models: A survey. *arXiv* (2024)
71. MirzaeiGhazi, S., Stenseke, J.: Responsibility before freedom: Closing the responsibility gaps for autonomous machines. *AI and Ethics*, 1–13 (2024)
72. Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D.C., Haber, N.: Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 599–627 (2025)
73. Morris, C.: Microsoft's new Bing AI chatbot is already insulting and gaslighting users. *Fast Company*. (2023). <https://www.fastcompany.com/90850277/bing-newchatgpt-ai-chatbot-insulting-gaslighting-users>
74. Morris, H.: Persons and punishment. *Monist* **47**(4), 475–501 (1968)
75. Nass, C., Moon, Y.: Machines and mindlessness: social responses to computers. *J. Soc. Issues* **56**(1), 81–103 (2000)
76. Nelkin, D.: *Making Sense of Freedom and Responsibility*. Oxford University Press, Oxford (2011)
77. Newall, P., Hayes, T., Weiss-Cohen, L., James, R.J., Byrne, C.A., Muggleton, N.: Could the provision of multiple game themes be a neglected gambling product structural characteristic? results of an online simulated gambling task. *Int. Gambl. Stud.* **25**(2), 366–385 (2025)
78. Nguyen, M., Bin, Y.S., Campbell, A.: Comparing online and offline self-disclosure: a systematic review. *Cyberpsychol. Behav. Soc. Netw.* **15**(2), 103–111 (2012). <https://doi.org/10.1089/cyber.2011.0277>
79. Novak, T.P., Hoffman, D.L.: Relationship journeys in the internet of things: a new framework for understanding interactions between consumers and smart objects. *J. Acad. Mark. Sci.* **47**, 216–237 (2019)
80. Oimann, A.-K.: The responsibility gap and LAWS: a critical mapping of the debate. *Philos. Technol.* **36**(1), 1–22 (2023)
81. Oliveira, M., Brands, J., Mashudi, J., Liefoghe, B., Hortensius, R.: Perceptions of artificial intelligence system's aptitude to judge morality and competence amidst the rise of Chatbots. *Cogn. Res. Princ. Implic.* **9**(1), 47 (2024)
82. Osman, M.: Psychological harm: implications for the scientific risk analysis and policy communities. *J. Risk Res.*, 1–6 (2025a)
83. Osman, M.: Psychological harm: what is it and how does it apply to consumer products with internet connectivity?. *J. Risk Res.*, 1–22 (2025b)
84. Parfit, D.: *On What Matters*, vol. 1. Oxford University Press, Oxford (2011)
85. Perales, J.C., Navas, J.F., Ruiz de Lara, C.M., Maldonado, A., Catena, A.: Causal learning in gambling disorder: Beyond the illusion of control. *J. Gambl. Stud.* **33**(2), 705–717 (2017)
86. Puntoni, S., Reczek, R.W., Giesler, M., Botti, S.: Consumers and artificial intelligence: an experiential perspective. *J. Mark.* **85**(1), 131–151 (2021)
87. Puranam, P.: The Pinocchio effect: AI and the risk of diminished human self-perceptions of agency. *J. Risk Res.* (2025)
88. Raff-Heinen, S., Brenk, S., Henneke, D.: Smart products 2.0: Redefining user risks with multimodal generative AI. *J. Risk Res.* (2025)
89. Ralph, K., Nunez, A.R.: Understanding the link between social skills and phone use. *J. Student Res.* **10**(3), 1–11 (2021)
90. Reviglio, U., Agosti, C.: Thinking outside the black-box: The case for "algorithmic sovereignty" in social media. *Soc. Media Soc.* **6**(2), 2056305120915613 (2020)
91. Ricoeur, P.: *Freedom and Nature: The Voluntary and the Involuntary* (E.V. Kohak, Trans.). Northwestern University Press, Evanston (1966)
92. Robbins, S.: Recommending ourselves to death: values in the age of algorithms. In: *Recommender systems: legal and ethical*

- issues, pp. 147–161. Springer International Publishing, Cham (2023)
93. Robbins, S., Blundell, I.: The case for an immediate ban on children's LLM chatbot use: applying lessons from two decades of social media harm. *AI Ethics* **6**(2), 171 (2026)
 94. Robbins, S., Blundell, I.: Losing our voice? generative AI and the degradation of human expression. *Minds Mach.* **36**(1), 2 (2026)
 95. Rubel, A., Jones, K.M.: Student privacy in learning analytics: An information ethics perspective. *Inf. Soc.* **32**(2), 143–159 (2016)
 96. Russell, P.: Strawson's way of naturalizing responsibility. *Ethics* **102**(2), 287–302 (1992)
 97. Saeidnia, H.R., Hashemi Fotami, S.G., Lund, B., Ghiasi, N.: Ethical considerations in artificial intelligence interventions for mental health and well-being: ensuring responsible implementation and impact. *Soc. Sci.* **13**(7), 381 (2024)
 98. Sahebi, S., Parvizi-Wayne, D.: LLM-“Friends” are hostile scaffolds in the age of loneliness. (2026)
 99. Santoni de Sio, F., Mecacci, G.: Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philos. Technol.* **34**(4), 1057–1084 (2021)
 100. Santoni de Sio, F., Van den Hoven, J.: Meaningful human control over autonomous systems: A philosophical account. *Front. Robot. AI* **5**, 323836 (2018)
 101. Scanlon, T.M.: *What We Owe to Each Other*. Harvard University Press, Cambridge (2000)
 102. Scanlon, T.M.: How I am not a Kantian. In: Parfit, D. (ed.) *On What Matters*, vol. 2, pp. 116–140. Oxford University Press, Oxford (2011)
 103. Schermer, B.W.: The limits of privacy in automated profiling and data mining. *Comput. Law Secur. Rev.* **27**(1), 45–52 (2011)
 104. Scholich, T., Barr, M., Stirman, S.W., Raj, S.: A comparison of responses from human therapists and large language model-based chatbots to assess therapeutic communication: mixed methods study. *JMIR Ment. Health* **12**(1), e69709 (2025)
 105. Schüll, N.D.: *Addiction by Design: Machine Gambling in Las Vegas*. Princeton University Press, Princeton (2025)
 106. Schuster, N., Lazar, S.: Attention, moral skill, and algorithmic recommendation. *Philos. Stud.* **182**(1), 159–184 (2025)
 107. Sharifzadeh, R.: ChatGPT as co-author? AI and research ethics. *Ethics Prog.* **15**(1), 155–173 (2024)
 108. Shavit, Y., et al.: Practices for governing agentic AI systems. *OpenAI* (2023)
 109. Shelby, R., Rismani, S., Henne, K., et al.: Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 723–741. Association for Computing Machinery (2023)
 110. Shockley, J.S., McGurn, W.C., Gunning, C., Graveley, E., Tillotson, D.: Effects of calculator use on arithmetic and conceptual skills of nursing students. *J. Nurs. Educ.* **28**(9), 402–405 (1989)
 111. Shoemaker, D.: Caring, identification, and agency. *Ethics* **114**(1), 88–118 (2003)
 112. Slattery, P., Saeri, A.K., Grundy, E.A., et al.: The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *arXiv* (2025). <https://arxiv.org/abs/2408.12622>
 113. Stark, M., Fins, J.J.: Engineering medical decisions: computer algorithms and the manipulation of choice. *Camb. Q. Healthc. Ethics* **22**(4), 373–381 (2013)
 114. Starke, C., Metikoš, L., Helberger, N., de Vreese, C.: Contesting personalized recommender systems: a cross-country analysis of user preferences. *Inf. Commun. Soc.* **28**(1), 41–60 (2025)
 115. Strawson, P.F.: *Freedom and Resentment and Other Essays*. Routledge, London (2008) [Original work published 1962]
 116. Strong, C., Ansons, T.: Between doubt and deference: Trust and harm in the risk society. *J. Risk Res.*, 184–189 (2025)
 117. Susser, D., Roessler, B., Nissenbaum, H.: Technology, autonomy, and manipulation. *Internet Policy Rev.* **8**(2), 1–22 (2019). <https://doi.org/10.14763/2019.2.1410>
 118. Synofzik, M., Vosgerau, G., Newen, A.: Beyond the comparator model: a multifactorial two-step account of agency. *Conscious. Cogn.* **17**(1), 219–239 (2008)
 119. Ta, V., Griffith, C., Boatfield, C., et al.: User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *J. Med. Internet Res.* **22**(3), e16235 (2020)
 120. Tufekci, Z.: Algorithmic harms beyond Facebook and Google: emergent challenges of computational agency. *Colo. Technol. Law J.* **13**(2), 203–218 (2015)
 121. Umbrello, S.: Meaningful human control over smart home systems: a value sensitive design approach. *Humana. Mente* **13**(37), 40–65 (2020)
 122. Valenzuela, A., Puntoni, S., Hoffman, D., et al.: How artificial intelligence constrains the human experience. *J. Assoc. Consum. Res.* **9**(3), 241–256 (2024)
 123. Victor, B.G., Goldkind, L.: The therapist in the machine: confronting AI's challenge to clinical social work. *J. Technol. Hum. Serv.* **43**(2), 73–81 (2025)
 124. Wada, K., Shibata, T., Saito, T., Sakamoto, K., Tanie, K.: Psychological and social effects of one year robot assisted activity on elderly people at a health service facility for the aged. In: *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pp. 2785–2790 (2005)
 125. Wallace, R.J.: *Responsibility and the Moral Sentiments*. Harvard University Press, Cambridge (1994)
 126. Wang, J., Tang, Y., Chen, X., Huang, S.S.: Overservice backfires: the impact of high robot attentiveness on tourist's discontinuous usage intentions. *Tour. Manag.* **113**, 105318 (2026)
 127. Watson, G.: Free agency. *J. Philos.* **72**(8), 205–220 (1975)
 128. Weidinger, L., Uesato, J., Rauh, M., et al.: Taxonomy of risks posed by language models. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229. Association for Computing Machinery (2022)
 129. Wolf, S.: Sanity and the metaphysics of responsibility. In: Schoeman, R. (ed.) *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, pp. 46–62. Cambridge University Press, Cambridge (1987)
 130. Wolf, S.: Hiking the range. In: Parfit, D. (ed.) *On What Matters*, vol. 2, pp. 33–58. Oxford University Press, Oxford (2011)
 131. Wood, S.: The deskilling debate, new technology and work organization. *Acta Sociol.* **30**(1), 3–24 (1987)
 132. Xue, G., He, Q., Lu, Z.L., Levin, I.P., Dong, Q., Bechara, A.: Agency modulates the lateral and medial prefrontal cortex responses in belief-based decision making. *PLoS ONE* **8**(6), e65274 (2013)
 133. Yoon, S.C., An, J.H., Choi, J.S., Chang, J.H., Jang, Y.J., Jeon, H.J.: Digital psychiatry with chatbot: recent advances and limitations. *Clin. Psychopharmacol. Neurosci.* **23**(4), 542 (2025)
 134. Zhang, R.W., Liang, X., Wu, S.H.: When chatbots fail: exploring user coping following a chatbots-induced service failure. *Inf. Technol. People* **37**(8), 175–195 (2024)
 135. Zuboff, S.: *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, New York (2019)