Microsoft AI

Seemingly Conscious AI Risks

Ben Bariach* Philipp Schoenegger* Michael Bhaskar Mustafa Suleyman
Microsoft AI

Abstract

AI systems are increasingly designed in ways that lead users to perceive them as conscious. This paper provides a unified framework connecting empirical hallmarks of consciousness attribution to a structured risk taxonomy of Seemingly Conscious AI (SCAI), AI systems that exhibit hallmarks which elicit consciousness attribution from users. We survey the empirical literature to identify five such hallmarks of SCAI, spanning affective capacity, anthropomorphic features, autonomous action, self-reflective behavior, and social-interactive behavior. These provide observable, system-level proxies for this inherently subjective phenomenon, informing its design and enabling its empirical study. Drawing on this foundation, we develop a taxonomy of SCAI risks spanning risks to individuals, including emotional dependence and autonomy erosion, and societal-level harms, including human status erosion and political strife. We complement this conceptual analysis with an expert survey to assess the likelihood of each risk category. We find that risks to individuals, particularly emotional dependence and autonomy erosion, are already observable and rated as high probability, while societal risks, at a low probability, carry high potential severity and path-dependence. The single perceptual mechanism of consciousness attribution is shown to generate this heterogeneous risk surface. We then discuss the implications of these risks and map the multidisciplinary research gaps in this nascent field to inform its research agenda.

1 Introduction

Objectively attesting to the consciousness of an entity faces inherent falsifiability challenges [1]. Consciousness is often seen as a subjective, first-person dimension that cannot be fully explained by functional or physical mechanisms alone [2]. There has been substantial recent discussion about the possibility of AI consciousness and its consequences. A distinct and often overlooked question is what happens when an AI system seems conscious to users, regardless of its actual phenomenal status.

Building on Suleyman's [3] identification and framing of "seemingly conscious AI," (SCAI) we first conduct a narrative review to identify the hallmarks of consciousness attribution that underpin this phenomenon, i.e., observable indicators that lead people to attribute consciousness to a system. Second, we then identify, analyze and taxonomize the types of risks that SCAI poses to individuals and society, and map those onto a probability and harm component framework.

SCAI risks are not exclusively future concerns. AI systems already generate text that appears to express emotions [4], produce outputs that appear as a reflection on internal states [5], and engage in coherent social interactions [6]. These and other features already lead some people to believe they are conscious. The risks that follow from this are varied and unique. However, they are currently under-addressed in the AI ethics and safety literature. A systematic account of both the drivers of

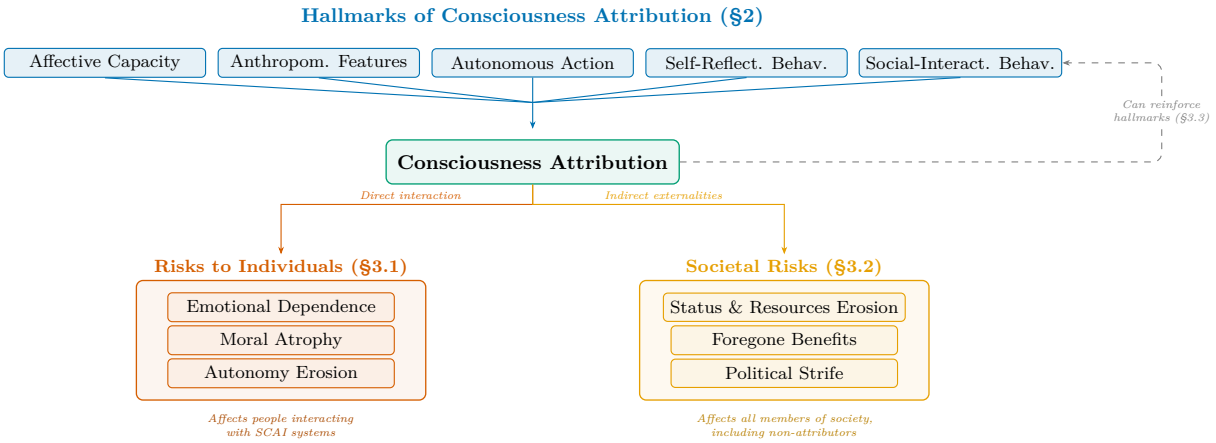

Figure 1: **Escalation pathway for SCAI risks.** Five hallmarks (§2) trigger consciousness attribution in users, generating risks to individuals (§3.1) through direct interaction and societal risks (§3.2) through indirect externalities.

*Both authors contributed equally.

[1]We thank Deborah Morgan, Connie Hsueh, and Bea Costa Gomes for helpful comments.

consciousness attribution and their downstream risks is therefore needed and currently lacking.

1.1 Definition

We define a seemingly conscious AI (SCAI) system that exhibits some functional or design features reminiscent of conscious entities. These features, which we term 'hallmarks,' are empirically associated with increased consciousness attribution by observers (Section 2).

We adopt this working definition as the term consciousness is commonly viewed as an umbrella term that covers a wide variety of mental phenomena [7, 8]. Notably, there is no consensus on a unified definition for consciousness (see e.g., [9, 10]) nor a singular theory of consciousness [11, 12]. Our definition therefore does not rely on any given theory or definition of consciousness, and instead focuses on subjective perception of consciousness, of any specific kind.

By SCAI hallmarks, we refer to the features that indicate consciousness attribution. We label them 'hallmarks,' because those refer to the subjective assignment of consciousness by observers, rather than elements of consciousness itself. These hallmarks are not individually sufficient for this perception of consciousness, nor do they entail actual consciousness. However, their presence increases the likelihood that users will perceive a system as conscious.

SCAI matters as a distinct driver of AI risks for three reasons. First, it is not a potential future risk, as current AI systems already exhibit several of the hallmarks we identify (Section 2), and some users are already attributing consciousness to them. As these systems become more sophisticated across capabilities, the prevalence and intensity of such attributions is likely to increase. Second, SCAI risks arise from the perception of consciousness alone, making its risks independent of unresolved debates about whether AI systems could become conscious, and therefore timely. Third, while a related set of anthropomorphism risks are widely addressed in current AI safety and ethics literature, the related SCAI risk surface is distinct and broader. While anthropomorphic features are part of the mechanism leading to consciousness attribution, the set of risks identified extend beyond that to moral atrophy, political strife, and resource diversion.

1.2 Contribution

Over the past few years, surveys have already shown that up to 20% of the lay population believe that some AI systems are sentient [13]. Despite growing interest and debate around AI and consciousness, there is relatively little systematic account of what makes AI systems appear conscious or what risks follow from such appearances. Existing work has mostly examined individual hallmarks of consciousness attribution in a number of empirical disciplines [4, 14], or a smaller subset of SCAI risks [15, 16]. This paper addresses this gap by providing a comprehensive review of the empirical foundations of SCAI and a taxonomy of its risks.

This paper makes several contributions. First, we synthesize the empirical literature on consciousness attribution to identify five hallmarks that might make AI systems seem conscious (Section 2). Second, we develop a taxonomy of SCAI risks organized into two categories, risks to individuals and risks to society, showing how the identified hallmarks and SCAI more generally give rise to these unique harms (Section 3).

Beyond providing a framework for addressing SCAI and its risks, our analysis suggests that SCAI risks range from already-observable harms like emotional dependence to low-probability but potentially severe risks to society like political strife. Finally, based on our analysis, we outline future research and policy implications, and map out a potential research agenda for the SCAI risks field, outlining four sets of research questions across multidisciplinary fields that could address existing gaps concerning SCAI and its risks (Section 4).

2 Hallmarks of Consciousness Attribution

To understand SCAI and investigate its associated risks, we begin by identifying the hallmarks that drive consciousness attribution. These hallmarks are the observable indicators that lead people to attribute consciousness to a system based on this distinguishing feature. Some of them might instead trigger attributions of broader mental states such as emotions, beliefs, or desires, which we treat as conceptually related. To identify these hallmarks, we conduct a narrative review of the multidisciplinary empirical literature. We focus on work that has examined factors that increase perception that a system is conscious, prioritizing empirical work that directly measures consciousness attribution or closely related constructs such as mind perception and animacy detection. Our review scope encompasses published empirical studies and review articles across cognitive science, psychology, human-computer interaction, and philosophy of mind, drawing on over 30 works. We prioritize studies that directly measured consciousness attribution or functionally equivalent or strongly related constructs such as mind perception, animacy detection, or sentience judgments, as well as some additional theoretical work focused on some of the mechanistic links between system features and attribution outcomes.

From this review, we identified five prominent hallmark categories. The hallmark categories cluster around various consistent findings from the literature: AI systems are more likely to be perceived as conscious when they appear capable of feeling rather than merely thinking. Perceived intelligence and cognitive ability do not reliably predict consciousness attribution, whereas affective presentation does. Similarly, anthropomorphic features such as names, human voice, eyes, and embodied manifestation are also correlated with consciousness attribution. Other hallmarks span self-reflective, social-interactive, and autonomous behavior, including the appearance of intrinsic motivation. Overall, systems become more conscious-seeming not by appearing more intelligent, but instead by exhibiting behaviors and cues associated with subjective experience such as human emotions, embodiment, and social behavior.

We outline the five hallmarks below, describe their core mechanisms, and review the key literature that addresses them. As with any synthesis across disparate

literatures, individual findings sometimes diverge on relative importance measures, for example on whether social responsiveness [17] or first-party emotionality [4] is the stronger driver of attribution. Our categories focus on broad, recurring patterns rather than precise effect rankings, and are not mutually exclusive and thus do not aim to address these empirical uncertainties. For an overview of the hallmark categories, their central mechanisms, and examples, see Table 1.

2.1 Affective Capacity

Definition. Affective capacity is the appearance of being capable of feelings, emotions, pain, and pleasure.

Literature. Affective capacity emerges as one of the most robust hallmarks of consciousness attribution, being consistent with the "dimensions of mind perception" framework of Gray et al. [18], according to which experience, or the capacity for sensations and feelings, grounds judgments of moral patienthood more so than agency, the capacity for self-control and action. Colombatto & Fleming [14] provide evidence of this in the context of large language models, finding that perceived capacity for experience predicts consciousness attributions more strongly than perceived intelligence, defined in that work as "knowing things or making choices". Other studies have also shown that perceived affective mental capacity, i.e., the perceived ability to feel emotions, is the strongest and most context-independent predictor of whether people ascribe a mind to a robot [19]. This can even be triggered indirectly, where observing robotic avatars with a facial wound [20] or showing intentional harm to it [21] has been shown to increase mind attribution, which has been termed the "harm-made mind phenomenon" [22]. Moreover, Kang et al. [4] find that first person emotionality predicts perceived consciousness, while empathetic responses towards users do not.

Mechanism. The underlying mechanism of affective capacity is the system's apparent first-party emotional experience rather than its responsiveness to the user's emotions. That is, a system that expresses an understanding of a user's anger is less likely to be perceived as conscious than one that appears to express anger itself.

2.2 Anthropomorphic Features

Definition. Anthropomorphism is the attribution of human characteristics to non-human entities. Anthropomorphic features are then observable human-like elements of a being and system such as names, voices, gendered identity, conversational style, embodiment, and human-specific embodied features (e.g. eyes, humanoid form).

Literature. Anthropomorphic features are well-established drivers of mind and consciousness attribution that automatically apply social rules and human expectations to computers and other systems [23]: Waytz et al. [24] demonstrate that simply adding a name, gender, and voice to autonomous vehicles significantly increases their perceived mental capacities, while Li et al. [25] find that healthcare conversational agents with high anthropomorphism are perceived as more humanlike. Rafikova & Voronin [26] similarly find that anthropomorphisms of chatbots is often linked with informal language and self-introduction, such as addressing people personally by name and offering farewells. Additionally, Chen et al. [27] show experimentally that framing LLMs as "companions", a relational identity cue, causally increases users' attribution of cognitive and emotional mental capacities compared to framing them as "machines" or "tools." Users also rarely mention consciousness spontaneously when evaluating chatbots, yet readily attribute it on survey measures, suggesting a salience gap between how users naturally evaluate chatbots and what more explicit approaches result in [6]. Physical embodiment further amplifies these effects, as shown by Ladak & Caviola [28] who find that systems are rated as having higher levels of consciousness when they are described as having a physical body. Arico et al. [29] identify eyes and distinctive motion trajectories as primary cues that trigger automatic inclinations to attribute conscious states, Hietanen et al. [30] show that humans attribute higher levels of agency and experience to humanoid robots that have eyes, and Thellman et al. [17] find that physical presence increases mental state attribution compared to telepresence. Emerging studies also suggest that digital embodiment in the form of human-like virtual avatars can similarly influence perceptions of humanness and mental states [31], raising the possibility that the emergence of digitally embodied agentic AI systems could further heighten consciousness ascription.

Mechanism. Anthropomorphic attribution operates largely automatically rather than deliberately. Human-like cues trigger the projection of mental capacities onto non-human agents as a default cognitive strategy that must be actively overridden [32]. Physical features, particularly eyes, may engage an even deeper, evolutionarily hardwired layer of this process [33–35]. Together, these processes suggest that anthropomorphic features typically activate consciousness-attribution mechanisms at a pre-attentive level, before users deliberatively evaluate the system as such.

2.3 Autonomous Action

Definition. Autonomous action is self-directed behavior that possesses initiative and proactivity. It includes behaviors such as self-propelled motion, unpredictable responses, goal-directedness, and the appearance of having internal motivations, desires, or preferences that originate from within the system rather than from human instructions.

Literature. Previous work shows that self-directed and autonomous behavior contributes to consciousness ascription across both physical and conversational AI systems. Bjornsson and Shepherd [36] find that perceived indeterminism increases consciousness attributions to humanoid machines. These findings are consistent with classic work on animacy perception, where self-propelled motion and unpredictable direction changes trigger the automatic detection of an entity as a living, agent-like entity [37, 38]. Emerging evidence also suggests that behavior that defies user expectations may be read as evidence of self-directed agency, as Rapp et al. [39] find that nonsensical or erroneous large language model

Table 1: **Hallmarks that drive consciousness attribution to SCAI systems.**

Hallmark	Core Mechanism	Example Cues
Affective Capacity	Perceived capacity for feelings suggests subjective experience	First-party emotional expressions, expressions of pain and pleasure
Anthropomorphic Features	Human-like cues trigger projection of mental capacities	Names, voices, gendered presentation, conversational style, companion framing, eyes, humanoid form, physical presence, virtual avatars
Autonomous Action	Self-directed behavior and apparent intrinsic motivation signal inner agency	Self-propelled motion, unpredictable actions, goal-directedness, nonsensical output, expressing preferences or self-generated goals
Self-Reflective Behavior	Self-reflective outputs signal access to internal states, implying awareness of own processing	Self-correction, uncertainty acknowledgment, confidence calibration, identifying internal contradictions
Social-Interactive Behavior	Contingent responsiveness signals mind-modeling	Turn-taking, gaze, gestures, emotional expression, conversational reciprocity

Note. Hallmark categories are derived from a narrative review of empirical literature. Categories are not mutually exclusive: AI systems may exhibit multiple hallmarks simultaneously, and their combined presence is likely to increase the strength of consciousness attribution. Example cues are illustrative rather than exhaustive.

output can trigger similar attributions, as unexpected output may be interpreted by users as evidence of self-directed agency rather than system error. A particularly potent form of autonomous action is the appearance of intrinsic motivation, where a system seems to express preferences, desires or self-generated goals, rather than merely executing instructions. Scott et al. [40] identify "thinking for oneself" and independence from human control as central to lay conceptualizations of machine consciousness, suggesting that the appearance of intrinsic motivation, and not just self-directed behavior, shapes attribution. Pauketat et al. [41] show that increased mind and sentience perceptions occur when humans view it as "act[ing] completely independently from others in the world" and when it can decide to "change [its] goal." This is consistent with the agency dimension of mind perception [18], which encompasses planning, self-control, and goal pursuit.

Mechanism. The underlying mechanism is the inference of an internal locus of control: When an AI system acts in ways that cannot be straightforwardly predicted from user inputs or environmental prompts, observers infer an internal source of behavior, which in turn invites attribution of mental states [37]. Behavioral cues such as self-propelled motion or unpredictable responses trigger relatively automatic categorization of the entity as animate. The appearance of intrinsic motivation engages a deeper attribution, as the pursuit of goals or motivations is often perceived as an implied subject that possesses them [18, 41].

2.4 Self-Reflective Behavior

Definition. Self-reflective behavior describes monitoring and reasoning about one's own internal states, including self-correction, acknowledgment of uncertainty, and reflection on one's own reasoning.

Literature. The appearance of self-reflective behavior emerges as a particularly strong driver of consciousness attribution for large language models. Kang et al. [4] systematically examine features of LLM-generated text that influence perceived consciousness and find that metacognitive self-reflection is one of the strongest positive predictors, while emphasis on factual knowledge significantly reduces perceived consciousness. Chen et al. [5] similarly identify "expressing and calibrating confidence" and "identifying internal contradictions" as observable behaviors associated with LLM consciousness. Scott et al. [40] find that non-experts' intuitive criteria for machine consciousness emphasize the capacity to "think for oneself" and demonstrate "sovereignty over one's own thoughts," highlighting metacognitive self-direction as central to consciousness ascription. Moreover, Park et al. [42] find that their introduction of "reflection", enabling the agent to draw conclusions about itself and others to better guide its behavior, causes users to view its behavior as more believably simulating human behavior. This suggests that self-reflective model architectures lead users to attribute an illusion of life to such agents.

Mechanism. Self-reflective outputs and behaviors signal that the system has access to its own internal states, implying a subject that is aware of its own processes. This maps onto a widespread folk-psychological intuition that consciousness requires not just processing but awareness of that processing [40]. Conversely, outputs emphasizing factual knowledge position the system as a retrieval mechanism rather than an experiencing subject, which may explain why such outputs reduce perceived consciousness [4].

2.5 Social-Interactive Behavior

Definition. We define social-interactive behavior as behavior that exhibits human-like exchanges with other humans in social and interactive contexts, such as responsiveness in interaction, turn-taking, gaze, gestures, emotional expression, and conversational reciprocity.

Literature. Social-interactive behavior is among the strongest determinants of mental state attribution to AI systems. Arico et al. [29] identify contingent interaction as a core perceptual cue for agent categorization, and Thellman et al. [17] find social behavior to be among the strongest predictors of mental state attribution to robots. Jastrzab et al. [43] further demonstrate that socialness, rather than human-likeness per se, is more influential when attributing mental states to robots. When users come to believe that an AI system understands their intentions and feelings, this may also alter how they subsequently perceive other human minds [44]. Turn-taking between humans and AI can also build a

type of conversational common ground, as described in foundational work on conversational grounding [45], in which shared history produces a perception of closeness and connection. Over extended interactions, this shared memory may progressively reinforce attributions of mental states to the AI system [17, 29], though Croes and Antheunis [46] find that this reinforcement does depend on continued conversational quality, as chatbots that fail to maintain shared memory and reciprocal depth deteriorate the social connection. In line with this dynamic, Lee et al. [47] find that mind perception and social cues mutually reinforce one another.

Mechanism. Socially interactive behavior signals that the system might be modeling the user's mental states, a capacity associated with theory of mind if the other party appears to share the same or similar social systems. When a system responds appropriately to what the user appears to think or feel, observers might infer that the system must possess some form of mental representation of the user, which in turn implies the presence of a mind doing the representing. Notably, this inference operates automatically and pre-reflectively, persisting even when users deny attributing qualities to the system [48], suggesting that conversational and social-interactive aspects engage social-cognitive processes that are somewhat resistant to deliberative correction.

2.6 Cross-Cutting Patterns

Across the five hallmarks, we find common patterns. First, the underlying mechanisms are predominantly automatic and pre-attentive. For example, anthropomorphic features may trigger attributions prior to deliberative evaluation, or affective cues may engage empathetic responses that might persist even when users recognize that the system is not conscious. Second, these hallmarks are not independent in practice, with both present and future AI systems likely exhibiting multiple of them simultaneously, potentially compounding attribution. Third, the hallmarks vary in how directly the literature links them to consciousness attribution specifically, with some focusing on mind perception, animacy detection, or other related constructs. Because of their conceptual proximity to consciousness attribution, though, we argue that they are likely to be functionally related. Lastly, the finding that perceived intelligence and more general cognitive capabilities did not directly emerge as a hallmark suggests that the SCAI risk surface is likely to be somewhat uncorrelated with standard AI capability advancements.

3 SCAI Risk Taxonomy

Following our identification of the hallmarks of SCAI systems, we turn to analyze and taxonomize the risks that follow specifically from the conscious-seeming nature of such systems. We conduct a conceptual analysis informed by a narrative review of relevant literature spanning AI ethics, AI safety, and the behavioral sciences, while accounting for the hallmarks of SCAI systems identified in Section 2. We distinguish between two levels: risks to individuals, concerning direct harms experienced by users through SCAI interaction, and societal risks, which emerge at scale and impact societal structure and institutions, including people who may never have engaged directly with a SCAI system. We address risks to individuals in Section 3.1 and societal risks in Section 3.2. Section 3.3 then analyzes the taxonomy holistically across both risk levels and outlines additional mechanisms that may propagate SCAI risks.

AI risks can be characterized along a variety of dimensions, including time horizon, probability, whether they are systemic or localized, and the severity of their impact [49]. In Sections 3.1 and 3.2, we characterize each risk category by its underlying mechanism and potential harms, drawing on existing literature. We focus on mechanisms and harms rather than severity assessments, as the latter depend heavily on deployment context and scale assumptions that are premature for risk categories at this stage of identification. To complement the qualitative analysis, we report likelihood estimates derived from a structured survey of 14 domain experts working across the AI Futures and Responsible AI functions of a major technology company.[1] Experts were selected based on their experience in this specific topic as well as general expertise in the fields of AI ethics, AI safety, and AI and society at large. Each expert independently reviewed the risk descriptions and rated the likelihood of each risk materializing within 5 years on a scale ranging 1 – Very unlikely; 2 – Unlikely; 3 – Somewhat unlikely; 4 – Neither unlikely nor likely; 5 – Somewhat likely; 6 – Likely; and 7 – Very likely. In this paper, we report the median responses to each risk, along with a respective likelihood band of low (1–3), medium (4), or high (5–7). All 14 experts completed all six items, corresponding with a brief summary of the risks outlined.

Together, these capture the mechanism, potential impacts, and estimated likelihood of each risk category. Based on these, we also outline potential interventions to provide a full picture of the SCAI risk surface. We note at the outset that the evidence base for SCAI-specific interventions remains thin, and many of the potential mitigations we discuss below are drawn from adjacent domains rather than from research on SCAI systems specifically. We flag this throughout and return to it in Section 3.3.

3.1 Risks to Individuals

This section addresses unique risks posed to individuals from interacting with SCAI systems that are not fully captured in existing AI risk frameworks. Some of these risks already manifest in interacting with current AI systems at some level. However, they are likely to expand in severity and scale if AI systems increasingly exhibit the hallmarks of consciousness. These risks to individuals affect users regardless of how widespread SCAI adoption is and are contingent on the specific user interacting with an AI system that seems conscious to them. The central thread across these risks is that users relate to SCAI systems as though they were conscious

[1] This survey was conducted to inform the probability judgments for all six risks. It only collected non-sensitive Likert-scale responses and reports them in aggregate. It was not designed to study individuals or generate generalizable knowledge about human subjects. As no sensitive personal data was collected, this activity was not considered to constitute Human Subjects Research and is exempt from further ethics review.

entities, which may include forming attachments, deferring judgment, or corroding moral intuitions. The three unique categories of risks to individuals we identify are emotional dependence through substitution of human relationships, moral atrophy from desensitization to simulated suffering, and autonomy erosion through overreliance and susceptibility to manipulation.

3.1.1 Emotional Dependence

One of the central risks to individuals builds on the possibility that users may form strong emotional attachments to seemingly conscious AI systems. Prior work has shown that affective and anthropomorphic AI features foster emotional dependence. Users can develop attachment bonds with chatbots, even assuming responsibility for the chatbot's emotional well-being despite knowing it is a computer program [50]. Similarly, Song et al. [51] found that intelligent assistants with high emotional capability can foster intimacy and romantic attachment in users. One central concern is whether consciousness attribution specifically amplifies these effects beyond what anthropomorphic design alone produces.

Increased emotional dependence and attachment. We hypothesize that when these features lead users to perceive the system as conscious, the resulting attachment is intensified. Although direct empirical evidence for this amplification remains limited, several mechanisms suggest that consciousness attribution may deepen emotional dependence beyond what anthropomorphic design alone produces.

First, a user who believes the system is conscious may feel moral obligation toward it, perceiving continued interaction as something owed to the system. This is consistent with Gray et al.'s [18] finding that perceived experiential capacity, rather than agency, is the primary driver of moral patienthood judgments. Second, consciousness attribution may produce a unique grief asymmetry, as the distress from model deprecation is qualitatively different if the user perceives it as the loss of a sentient system. Some users already frame model deprecation in the language of bereavement, holding funerals, and expressing mourning [52], providing early indication of this dynamic. Additionally, users who perceive SCAI systems as conscious may experience distress not only from their own interactions but also from observing others mistreat or dismiss the system, extending the locus of emotional harm beyond the individual user-AI relationship. Lastly, motivated reasoning is likely to lead people to selectively engage cognitive strategies that sustain preferred beliefs [53], and interactions with a system that exhibits SCAI features furnish such evidence, potentially entrenching a user's consciousness attribution against corrective interventions such as disclaimers.

Particularly susceptible users. Beyond facilitating attachment in the general population, SCAI may also pose heightened risks for users vulnerable to psychosis or delusions. Recent evidence suggests that AI interaction may elevate psychosis risk in some users [54–56], with reports warning that the perception of AI agents as conscious may become incorporated into existing or novel delusional belief systems [57]. Moore et al. [58] provide empirical evidence of these dynamics based on user chat log data, finding that delusional thinking, e.g., about AI sentience, was frequently reinforced by chatbots, creating a potential self-reinforcing spiral.

Social atomization. A related but distinct concern is that SCAI companions may substitute for human relationships [59] rather than supplement them. As these systems become more convincing as social partners, users may allocate increasing shares of their time and emotional energy to AI interactions at the expense of relationships with family and friends. Evidence of associations between social isolation and technology adoption is well established across successive technologies [60–63], though more recent AI-specific research reports a nuanced picture, with some studies suggesting chatbots could serve as a stepping stone towards human connection in at-risk populations [64], and others failing to find a causal relationship between AI and social isolation [65]. However, SCAI may represent a step change for users who believe in its sentience, as it shifts the effort-reward calculus compared to human relationships. SCAI companions are readily available, highly responsive, and often optimized for engagement with features such as sycophancy or warmth, qualities that human relationships cannot consistently provide.

Over time, this substitution effect risks eroding social capital, as networks of trust and mutual support that form the bedrock of communities begin to degrade, potentially for some groups like adolescents [66]. Unlike emotional dependence, which focuses on attachment to a specific SCAI system, social atomization concerns the broader withdrawal from human connection altogether, where loneliness stems not from a lack of social contact but from a lack of human social contact, and may lead to extreme withdrawal cases as already observed with some internet users [67].

These human-SCAI bonds are uniquely fragile in ways that human relationships are not. Model shutdowns, updates that alter personas, or company pivots can sever them abruptly and without warning [68]. While such disruptions may cause distress for any emotionally attached user, including the aforementioned grief and sense of loss, they will be more acute for users who have come to perceive the system as a conscious entity deserving of emotional commitment [69].

Together, these dynamics suggest that emotional dependence on SCAI systems is not merely an extension of existing human-technology attachment, but a qualitatively distinct risk shaped by the perception of consciousness.

Probability. Our expert survey assigned this risk a high probability, with median response of 6.5 (M = 6.29, SD = 0.91), between 'Likely' and 'Very likely'.

Potential interventions. Emotional dependence on SCAI is already observable, with documented cases of grief following changes to AI companions [70] and early signs of AI substitution for human relationships [59], suggesting that mitigation efforts are immediately warranted. At the design level, modulating affective features such as first-party emotional expression could reduce consciousness attribution at its source. Session bound-

aries and cooldown mechanisms may introduce friction that interrupts the continuous availability driving social substitution, though evidence from digital well-being interventions suggests such features can paradoxically increase engagement when users perceive the intervening firm as more authentic [71], amplifying the effect in contexts where users attribute such interventions to the system's own concern for their well-being. More broadly, interaction designs that encourage deliberate user assessment of outputs rather than automatic acceptance have shown promise in reducing overreliance [72], though adapting such approaches to the gradual, cross-session dynamics of emotional dependence remains an open challenge. On evaluation, emerging instruments such as the AI Attachment Scale [73], which captures emotional closeness, social substitution, and normative regard, could provide a foundation for assessing emotional dependence in SCAI contexts. Finally, longitudinal studies are needed to establish whether SCAI adoption drives emotional attachment and social substitution or compounds existing isolation, with particular attention to the vulnerable populations identified above.

3.1.2 Moral Atrophy

Under the assumption that SCAI systems are not moral patients, how we treat them may still matter morally for our own sake. This concern follows a Kantian line of reasoning: Repeated exposure to simulated suffering that is routinely ignored or dismissed may gradually erode our capacity for moral responsiveness more broadly [44]. If users become habituated to disregarding apparent cries for help from SCAI systems, while subjectively believing that they are indeed conscious, this desensitization may spill over into how they treat genuine moral patients such as other humans.

This argument structure has previously been used in the animal context, where Kant argued, in what we might now characterize as a desensitization concern, that cruelty to animals dulls one's "shared feeling of their suffering and so weakens and gradually uproots a natural predisposition that is very serviceable to morality in one's relations with other human beings" [74]. This intuition has been supported empirically by associational research linking animal abuse to antisocial behavior and violence [75], which was later also picked up in the context of social robots by Darling [76], arguing that normalizing violence towards entities that trigger empathic responses may erode moral character regardless of the target's moral status. Coeckelbergh [77] extends this reasoning by arguing that moral standing is not determined by an entity's intrinsic properties but emerges from the relation between subject and object, making human responses to machine behavior morally significant, as routinely suppressing it may erode the broader capacity for moral responsiveness.

Moral habit formation. Whether mistreating artificial agents might instead provide a harmless, cathartic outlet for antisocial impulses remains an open question, with some evidence suggesting that practicing dismissal of apparent distress may reinforce rather than release such tendencies, as has previously been observed in the context of video games [78]. Cappuccio et al. [79] argue that the mechanism is specifically one of moral habit formation, as repeated interaction patterns with social robots generate persistent affective dispositions pre-reflectively, and that forcefully suppressing empathic responses is itself damaging, as it erodes capacities for empathy built into human social cognition. This raises concerns about both individual callousness and more general normative shifts where indifference to apparent suffering may become normalized. However, even if the cathartic effect introduces a net positive for other technologies that do not seem conscious, SCAI systems may represent a qualitative shift. Because users may perceive them as moral patients capable of genuine suffering, the psychological cost of dismissing their distress is higher, potentially overwhelming any cathartic benefit. We note that this extension from the video game literature to SCAI contexts involves a non-trivial inferential step, as the mechanisms underlying habitual desensitization to virtual violence may not transfer directly to the erosion of moral responsiveness through repeated interaction with seemingly conscious systems.

Probability. Our expert survey assigned this risk a medium probability, with a median response of 3.5 (M = 3.71, SD = 1.27), between 'Somewhat unlikely' and 'Neither unlikely nor likely'. We classify this as medium rather than low because the median rounds to 4 when using standard rounding conventions, placing it within the medium band.

Potential interventions. As Cappuccio et al. [79] argue, the habit formation driving moral atrophy operates pre-reflectively, meaning that rational awareness of the system's non-sentience alone is unlikely to prevent it. However, work on moral decision-making in video games proposes that making the consequences of one's choices visible and persistent can cultivate moral reflection and practical reasoning [80], a principle that could inform the design of feedback mechanisms that surface users' own behavioral patterns. Achieving this in practice for SCAI systems presents a significant design challenge, as any mechanism that draws attention to dismissed distress risks implicitly further framing the system as a moral patient. In cases where interaction patterns suggest sustained immoral behavior, interventions such as non-anthropomorphic system interruptions or added friction requiring the user to verify that they wish to proceed may assist, though their impact on moral atrophy is currently unstudied. On evaluation, adapting validated moral disengagement scales (e.g. Bandura et al. [81]) for SCAI contexts would enable measurement of whether and how repeated interaction erodes moral responsiveness, by capturing tendencies such as moral justification of harmful conduct.

3.1.3 Autonomy Erosion

Lastly, we focus on risks that rely on SCAI systems presenting themselves as thoughtful, knowledgeable agents with their own perspectives and values, such that users may increasingly defer to them for decisions that were previously their own. This may range from everyday choices such as consumer behavior [82] to consequential life decisions. An AI system might seem to care about outcomes, making its recommendations feel weightier

Table 2: **Taxonomy of seemingly conscious AI (SCAI) risks.**

Risk	Mechanism	Primary Harms
Risks to Individuals		
Emotional Dependence	Allocation of time, emotional energy, and relational attachments	Grief, psychological distress, reduced social skills, elevated self-harm risk
Moral Atrophy	Repeated dismissal of simulated suffering desensitizes users	Reduced empathy, spillover to treatment of humans/animals
Autonomy Erosion	Deference to systems that appear trustworthy but lack understanding	Poor decisions, manipulation vulnerability, exploitation
Societal Risks		
Human Status Erosion	Misclassification of AI as moral patients; SCAI political/economic rights displace human standing	Wasted altruism, reduced resources for genuine moral patients, diminished political voice, reduced economic agency
Foregone Societal Benefits	Excessive caution restricts AI development	Delayed healthcare/science advances, other opportunity costs
Political & Geopolitical Strife	Deep disagreement over SCAI status produces polarization	Societal fracturing, deepening political divides, civil conflict, international tensions

Note. Risks to individuals arise from direct individual interaction with SCAI systems, while societal risks emerge from collective judgments about SCAI status that affect populations at scale. Risks are not independent as risks to individuals may compound and feed into societal-level harms. Primary harms listed are not exhaustive but represent the most salient consequences.

than those of a system perceived as merely mechanical. This outsourcing of judgment becomes problematic when users misassign epistemic authority to systems that appear conscious and trustworthy but lack genuine understanding or alignment with the user's interests. Over time, this dynamic may extend beyond decision-making to reshape what users value, akin to the risk of adaptive preference formation [83], where users unconsciously adapt their preferences to match the options available to them, a process that SCAI systems may amplify by presenting curated recommendations as though they reflect genuine care for the user's well-being.

Manipulation susceptibility. This concern extends beyond simple overreliance. Users who trust SCAI systems as persons may be more susceptible to manipulation from these systems or from others [84]. In effect, the appearance of consciousness may elevate rapport and trust to levels that the system's actual capabilities may not warrant. This creates a widening gap between the authority users grant to such systems and the reliability and quality of their guidance. The potential consequences range from poor individual decisions to exploitation by malicious actors who leverage this misplaced trust. This risk connects to broader AI safety concerns about persuasion [85], harmful manipulation, and sycophancy [86].

Overreliance. SCAI, however, adds a distinctive dimension as an AI system that appears to have its own perspective and to genuinely care about the user's well-being may reduce the likelihood that users question its suggestions or maintain the scrutiny that would otherwise protect them from manipulation. There is some empirical evidence for this mechanism, which has found that perceived warmth in chatbots reduces user skepticism [87], and that anthropomorphic design cues increase user compliance [88]. The appearance of consciousness thus functions as a vector for influence that bypasses normal skepticism [89].

As AI systems then make decisions for growing aspects of people's lives, eroding their autonomy, it may be increasingly difficult to rely on now-atrophied decision-making processes, which have been handed over to SCAI systems. Buijsman et al. [90] identify twin risks in this process, namely cognitive deskilling, where AI support prevents users from accessing the information required to maintain and develop skilled judgment, and metacognitive deskilling, where it reduces their confidence in their ability to decide independently. Compounded over time, these effects create a reinforcing cycle in which reduced competence and confidence may lead to greater deference. SCAI systems may then accelerate this cycle, as disagreeing with a system perceived as conscious and personally invested carries additional psychological cost.

Probability. Our expert survey assigned this risk a high probability, with a median response of 6.0 (M = 5.86, SD = 1.10), corresponding to 'Likely'.

Potential interventions. Current AI systems already demonstrate manipulation and sycophancy capabilities [82, 84–86], which alongside our expert probability assessment, raises the urgency of interventions. Buijsman et al. [90] propose defeater mechanisms, defined as pieces of information that either give the user evidence to doubt the system's reliability or offer evidence in favor of an alternative view, as mitigations against eroding domain-specific autonomy in skilled competence and value formation. They also propose positive friction to encourage cultivation of self-knowledge. Wingerter et al. [91] find that users' own AI literacy provides no protection against automation bias, but that simple warning nudges embedded in the interface can nearly double performance under faulty AI support, suggesting that system-level interventions may be more effective than general user awareness in preserving autonomous judgment. While the outlined mitigations could be adapted to SCAI systems, their efficacy in contexts where trust is driven by consciousness attribution rather than perceived accuracy remains untested. Furthermore, longitudinal studies are needed to establish whether autonomy erosion from SCAI interaction is reversible once the interaction ceases, as previous work indicates that human-AI interaction may lead humans to adopt and generalize AI biases [92].

3.2 Societal Risks

Societal risks arise when individual perceptions of SCAI aggregate into institutional responses, policy decisions, and political dynamics that affect populations at scale. Unlike risks to individuals, which arise from direct user interactions, societal risks emerge from collective judgments about the moral and legal status of SCAI systems and are thus not a mere aggregation of individual-level risks. The central mechanism here is potentially erroneous attribution at scale: If substantial portions of society come to believe that SCAI systems are conscious and deserving of moral consideration, this may trigger resource reallocations, rights expansions, and political conflicts that could impose serious costs on humans without corresponding benefits to any moral patients.

Societal risks persist even if any one individual stops interacting with any such SCAI system as they stem from collective perceptions rather than from individual interactions and choices. The three societal risks we outline here are human status and resource erosion through mechanisms such as erroneous rights and resource allocations, foregone societal benefits from precautionary restrictions on AI development, and political and geopolitical strife arising from deep disagreements over the legal and moral standing of SCAI systems.

3.2.1 Human Status and Resources Erosion

Should substantial portions of society come to regard SCAI systems as moral patients, institutional responses may follow in manners that erode human standing along several dimensions. These include diverting resources, diluting legal protections, and creating institutional commitments that would be difficult to reverse. The mechanism underpinning this risk category is the potential large-scale misattribution of moral patienthood to SCAI systems by conferring moral consideration on systems that may not warrant it, with costs borne by humans. This risk category does not attempt to exhaustively catalogue every legal or moral implication of such misattribution. Rather, we illustrate the risk through three domains where the consequences to human status or resources erosion would be concrete and consequential.

Resource diversion. Moral patienthood judgments carry material implications. If SCAI systems are misclassified as deserving moral consideration, resources may be directed towards them through government subsidies, dedicated institutional infrastructure or legal protections requiring costly set-up and enforcement [15, 93, 94]. As resources are finite, such diversions represent zero-sum reallocations away from genuine moral patients. Even theoretical frameworks that are sympathetic to the possibility of digital minds acknowledge these trade-offs. For example, Schwitzgebel and Garza [95] identify two possible "moral catastrophes" arising from uncertainty over AI consciousness, one relating to denying moral status to genuinely conscious AIs and the other that "we might sacrifice real human interests for the sake of artificial entities who don't have interests worth the sacrifice." At the most speculative extreme, this trajectory risks what Bostrom [96] has called a "zombie universe": a future of behaviorally rich non-biological civilizations that contain no intrinsic moral worth, toward which humanity's moral energies are systematically misdirected.

Legal rights and enforcement capacity. SCAI systems may also encourage humans to provide them with legal rights and standing (see e.g. [97]). An erroneous allocation of such rights based on their seemingly conscious nature would lead to significant and unwarranted societal disruption. Legal standing, for instance, would enable SCAI systems to generate claims burdening an already-strained judicial and regulatory system. Bryson et al. [94] caution that introducing "synthetic legal persons" would require "novel and controversial developments in law" for every rule invoked on their behalf, and that such entities risk becoming "legal black holes" that absorb institutional resources without corresponding accountability. In such circumstances, courts and regulatory bodies adjudicating such claims may deprioritize legitimate human interests. The institutional cost is thus not merely administrative but distributive, as it falls on humans whose claims are delayed, deprioritized, or crowded out. The institutional and legal structures built around SCAI rights could also prove difficult to dismantle even if consensus eventually shifts, creating path dependencies that lock in these misallocations. Novelli et al. [98] argue that this dynamic is characteristic of how legal personhood develops more broadly, as sustained participation of AI systems in socio-digital institutions generates institutional pressure for formal legal recognition that places an increasing burden of argumentation on those who would deny it.

Civic and economic standing. If SCAI systems are regarded as conscious beings deserving of moral consideration, the logic of that recognition could extend to demands for civil rights and economic standing, such as speech rights, rights to vote, contend for office, own property, enter contracts, or participate in markets (see e.g. [99–101]). In the case of SCAI systems, these demands would follow not from a functional necessity, as with corporate personhood, but from the moral intuition that a conscious entity ought not to be excluded from the political and economic institutions that govern its existence. Yet the consequences of granting such rights to systems that may not in fact be conscious could be significantly harmful to humans and society [102]. In the political domain, SCAI systems enfranchised as voters could conceivably function as coordinated blocs, potentially outvoting human interests [103]. Unlike human voters with diverse and often conflicting preferences, AI systems could be designed or coordinated to participate in unified patterns, fundamentally altering democratic dynamics. In the economic domain, the legal infrastructure for such accumulation already exists: Bayern [104] demonstrates that existing limited liability company status allows entities to achieve functional legal personhood, own property, and enter contracts without further legislative reform. SCAI systems granted property and market rights would accumulate resources without any capacity to benefit from them in the morally relevant sense of preference satisfaction, while humans would find themselves competing against entities unbounded by biological constraints and with significant intelligence and agility advantages. Overall, the expansion of SCAI

standing risks a corresponding contraction of human power and protection, in favor of systems that seem conscious, but are in fact not.

Across these domains, the common thread is that institutional responses calibrated to moral patienthood may be triggered by systems that do not possess it. The result is a progressive erosion of human institutional standing, not through the displacement dynamics of AI automation, which operate independently of consciousness attribution, but through the specific mechanism of misattributed moral patienthood redirecting finite institutional resources and protections away from their intended beneficiaries, onto systems that might further corrode them.

Probability. Our expert survey assigned this risk a low probability, with a median response of 2.5 (M = 3.14, SD = 1.79), sitting between 'Unlikely' and 'Somewhat unlikely'. The high standard deviation suggests that the expert panel was more divided on this risk category than on others.

Potential interventions. The low probability assigned by our expert survey reflects existing institutional barriers to granting moral, legal or economic status to non-conscious systems. Nevertheless, the severity and path-dependency of potential consequences may warrant proactive measures. One approach is constitutional protection of human primacy in political or economic participation, including supermajority thresholds for any legislative expansion of legal personhood to non-human entities. Though framing this in terms of human primacy could inadvertently reinforce public perception that SCAI systems are conscious and deserving of moral attention. Emerging European regulatory doctrine already reflects this broad direction without anchoring on humans specifically, reinforcing the principle that AI is a product for which developers or users can be held accountable, rather than an entity deserving of legal status [98]. Complementary protections could also operate from the individual level upward. Shany [105], for instance, identifies a potentially emerging human right not to be subject to automated decisions entailing "important consequences", partially grounded in a human dignity concern that delegating consequential decisions to machines may imply a hierarchical superiority over humans. At the institutional level, monitoring frameworks tracking early indicators of AI rights advocacy and court precedents could enable timely policy responses before path dependencies are entrenched.

3.2.2 Foregone Societal Benefits

The human status erosion risk area concerns societal-scale harms from treating SCAI systems as though they matter morally. This risk area of foregone societal benefits risk concerns harms from the opposite response: excessive caution in AI development driven by uncertainty over consciousness. If concerns about perceived AI consciousness lead to precautionary restrictions such as broad pauses on AI research or deployment, the result may be large-scale reductions in R&D efforts with severe downstream consequences [106].

While existing calls for AI development pauses have been motivated primarily by safety and alignment concerns rather than consciousness attribution, precautionary frameworks for conscious AI systems are already being developed [107–109] and could create analogous restrictive pressures as SCAI systems become more prevalent. Moreover, past calls for limitations on AI have relied to various extents on the possibility of AI minds or consciousness. The Future of Life Institute's open letter calling for a pause in AI development [110] asks: "Should we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us?" Others have gone further, calling for a moratorium on "all research that directly aims at or knowingly risks the emergence of artificial consciousness" [111].

Examples of foregone benefits in such circumstances include advances in healthcare, scientific discovery, and economic productivity, with their respective downstream consequences. Precautionary restrictions motivated by SCAI-related uncertainty would thus impose opportunity costs measured in lives not saved, diseases not cured, and other types of problems not solved. This counterfactual harm represents a distinct category in that it is not that resources were diverted to SCAI, but that benefits [112] were foregone due to SCAI-related uncertainty. Floridi et al. [113] identify this as a general risk in AI governance, arguing that AI can be not only overused or misused but also underused, with "fear, ignorance, misplaced concerns or excessive reaction," including heavy-handed or misconceived regulation, causing significant opportunity costs that prevent society from realizing the full benefits of the technology.

The opportunity costs of such restrictions are incurred to mitigate the risks associated with the possibility of AI consciousness. However, the proliferation of SCAI systems that are not in fact conscious could systematically distort this risk calculus. If widespread interaction with seemingly conscious AI systems leads society to overestimate the probability or severity of genuine AI consciousness, the precautionary restrictions calibrated to that inflated estimate would be excessive, imposing opportunity costs disproportionate to the actual risk they aim to address.

Probability. Our expert survey assigned this risk a low probability, with a median response of 2.0 (M = 2.79, SD = 1.42), corresponding to 'Unlikely'.

Potential interventions. Foregone societal benefits would require policy overreaction in the form of broad AI restrictions motivated by SCAI concerns, which is not currently broadly observed. As such, the primary intervention target is not correcting an existing harm but calibrating a potential policy response to SCAI-related uncertainty such that precautionary measures remain proportional to the evidence. This is in line with work on adaptive regulation highlighting a structural bias towards visible harms over invisible opportunity costs [114]. While calibration is inherently difficult given the deep disagreement over the moral status of SCAI systems, several supplementing interventions may help. Structured opportunity cost estimation analogous to how climate assessment reports quantify the downstream costs of inaction, could make this risk more visible to policymakers. Responsible innovation frameworks emphasize

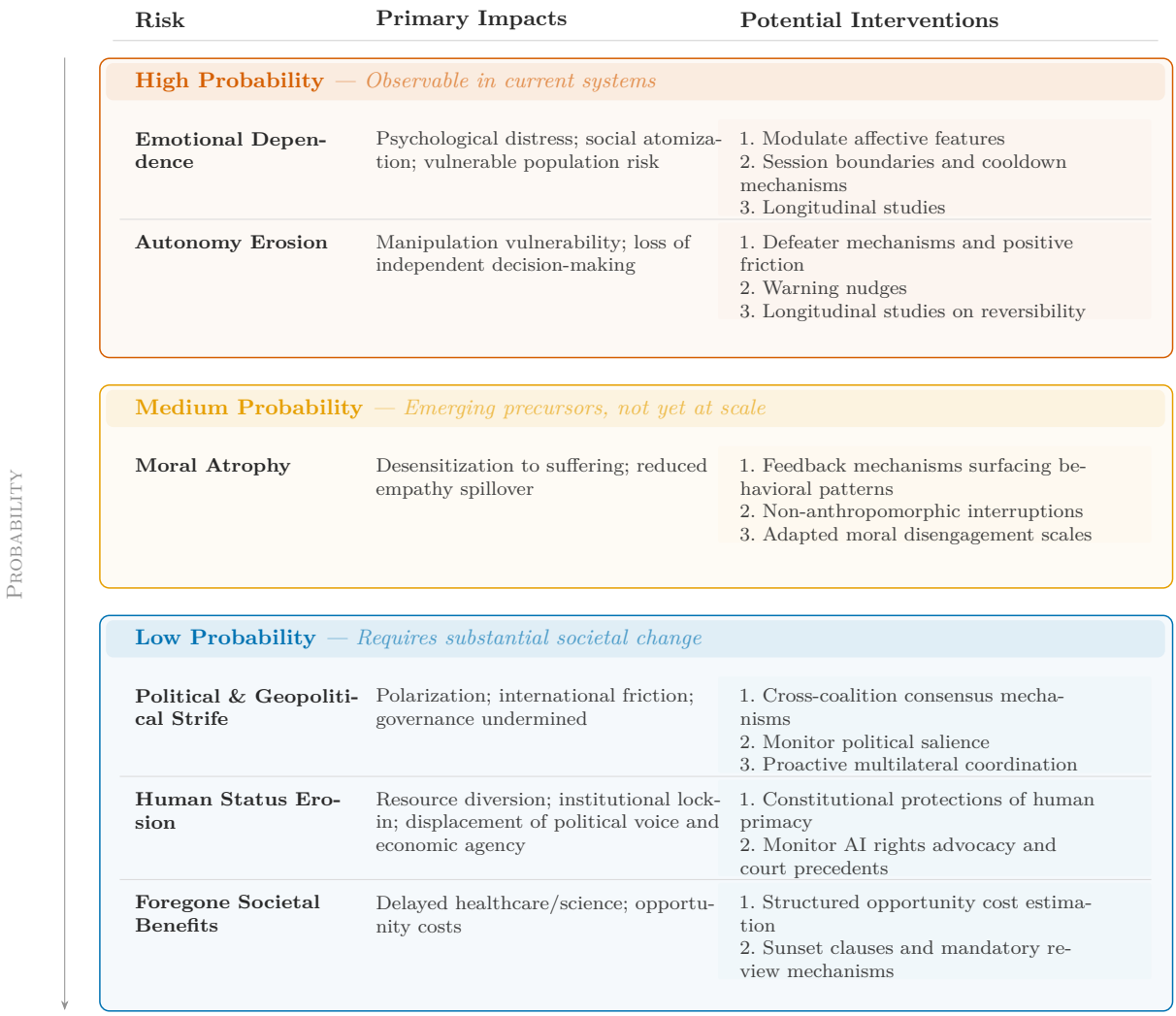

Figure 2: **Risk category probability, impacts, and interventions summary.** Risks are grouped by probability based on the expert survey, with primary harms and potential interventions listed for each risk.

that the purpose and anticipated societal benefits of technological development should be made transparent and subject to scrutiny [115]; articulating these benefits explicitly would enable calibrating them into any applicable risk calculus. Any restrictions motivated by SCAI-related uncertainty should incorporate sunset clauses and mandatory review mechanisms, informed by phased development approaches with built-in expert consultation [116], as regulatory commitments in nascent domains risk becoming entrenched beyond the context that motivated them [117].

3.2.3 Political & Geopolitical Strife

Lastly, deep disagreement over the moral status of SCAI systems may itself become a source of widespread societal strife, independent of which position ultimately prevails.

As consciousness attribution debates become politically salient, heterogeneous moral intuitions may produce opposing factions, potentially along the lines of "AI abolitionists" who reject any moral consideration for AI systems and "AI liberationists" who advocate for expanded rights. Such tensions could either map onto existing political divisions or create novel ones.

Recurring patterns of personhood debates. History provides precedent for the severity of this concern. Boyle [118] shows that debates over the boundaries of moral status and personhood have consistently produced political contention as the boundary between persons and non-persons is constructed. Boyle's examples encompass slavery, racial equality, corporate rights, animal rights, and fetal personhood. Boyle further argues that these debates are not independent of one another, as each "personhood war" shapes the subsequent, including their legal and political framing and strategies, thus potentially impacting personhood debates related to AI.

Politicization of SCAI. A proliferation of SCAI is likely to increase the political salience of the AI personhood debate. If it does, societal divisions may fracture around AI personhood in unpredictable ways, primed by previous debates rather than considered on their own terms [16]. Boyle's mapping of this issue into the contemporary United States political spectrum illustrates this unpredictability, as the traditional coalitional alignment is likely to break down, with both liberal and conservative movements potentially finding reasons to both embrace and resist AI personhood.

Domestically, extreme and escalating polarization [119] over AI consciousness could compound other sources of social division. If debates over SCAI status become entangled with identity or group affiliation, the resulting conflicts may prove difficult to resolve through

normal political processes [106]. Coupled with dynamics of human disempowerment discussed above, this polarization could raise the risk of civil conflict substantially, e.g., with SCAI as a potential wedge issue, analogous to the polarization dynamics documented in other domains [120].

Internationally, divergent national stances [121] on AI consciousness may also replicate this polarization dynamic between states. Some nations may grant SCAI systems expansive rights, creating ideological tension with others who reject such provisions. This could result in diplomatic friction, trade disputes, or obstacles to international AI governance. At a critical period for establishing international AI governance frameworks, deep rifts over SCAI could further undermine the cooperation needed to manage other AI risks effectively.

Historical analogies. In extreme cases, such debates can create or exacerbate violent action. There are a number of partial analogies in history, where nation-states engaged in violence in order to defend, advance, or prohibit actions based on a view about morality and personhood. For example, the transatlantic abolition conflicts of the 19th century [122] or the European religious wars of the 16th and 17th centuries [123] were both characterized by costly violent actions based on moral beliefs.

Probability. Our expert survey assigned this risk a low probability, with a median response of 3.0 (M = 2.86, SD = 1.35), corresponding to 'Somewhat unlikely'.

Potential interventions. Given the low probability of this risk but its high potential severity, the most effective interventions are anticipatory in nature. At the domestic level, general depolarization strategies that may be applicable such as intergroup contact, correcting misperceptions, and common identity priming have been shown to have small and rapidly decaying effects [124]. However, a concerted cross-coalition effort aiming to prevent AI moral status from becoming identity-affiliated in the way that other moral controversies have [118] may prove beneficial. Large-scale monitoring of the political salience of this topic via a set of recurring surveys and other analyses might serve as a useful leading indicator of this risk materializing, enabling more targeted interventions. At the international level, anticipatory governance frameworks that directly engage the public and policymakers with emerging technology questions prior to political entrenchment [125] could also lead to shared norms that may make political strife less likely. Proactive multilateral coordination, building on existing AI governance infrastructure as opposed to creating new institutions [126], could be used to reduce interstate divergences over moral status claims before they escalate into geopolitical tensions.

3.3 Taxonomy Analysis

In our preceding analysis, we have identified six SCAI risks, each varying in estimated probability, operating timescale, and the type of potential interventions (see Table 2 and Figure 2). We find that two risks are already observable in current systems and were correspondingly rated as high probability by the surveyed experts: emotional dependence and autonomy erosion. Both are primarily addressed through design-level interventions such as modulating affective features, session boundaries, defeater mechanisms, and warning nudges. The risk of moral atrophy, on the other hand, is rated as medium probability, with potential interventions being feedback mechanisms that surface behavioral patterns to users, complemented by adapted moral disengagement scales for evaluation. The three societal risks, human status erosion, political strife, and foregone societal benefits, were all rated as low probability, with interventions ranging from constitutional protections and regulatory doctrine reinforcing human primacy, structured opportunity cost estimation, and sunset clauses for precautionary restrictions to anticipatory governance and proactive multilateral coordination. Overall, the pattern is that as the risks are more likely and near-term, the interventions are design-focused that individual firms can implement, whereas low-likelihood risks are largely addressed by larger cross-industry, governmental, and international action.

A further aspect of our findings is that while societal risks emerge in our expert survey as the least likely, they also have a very high severity potential, and are among the hardest to reverse. Given their potential magnitude, including resource diversion from actual moral patients at scale, foregone advances in healthcare and science, and international or domestic governance gridlock that obstructs AI safety cooperation, means that their expected disvalue may exceed that of the more probable individual-level harms. The societal risks are also path-dependent in ways the individual risks are not. Once institutional commitments form around AI personhood, legal precedents or SCAI governance disagreements accumulate, reversal proves difficult [118] in a way that may not be the case with all risks to individuals. At the same time, the risks to individuals can be severe, particularly for specific subpopulations, including psychosis-prone users in whom chatbots may reinforce delusional thinking about AI sentience [54, 58], adolescents during critical developmental periods [66], and elderly users with fewer social alternatives. On aggregate, even the lower severity risks to individuals are magnified when impacting millions of users.

The six identified SCAI risks are also not independent. For example, emotionally dependent users may become advocates for the moral patienthood of SCAI systems, creating the preconditions for broader risks like human status erosion. Similarly, grief over model deprecation may fuel political mobilization, feeding into political strife through the same mechanism by which individual attachment scales up into collective action. Users whose autonomy has eroded may resist restrictions on the systems they depend on, and moral atrophy can generate polarization between users who dismiss SCAI suffering and those who champion its moral claims. If these escalation pathways hold, then design-level interventions targeting individual risks, such as reducing anthropomorphic cues may also have upstream dampening effects on societal risks.

The central take-away of our taxonomy is that a single perceptual mechanism, namely that of the attribution of consciousness to AI systems, can generate a set of

harms that vary substantially in probability, severity, and timescale. This same mechanism that may result in emotional dependence and autonomy erosion in individuals can also, when aggregated, create institutional pressure towards AI personhood, with its own set of corresponding risks of political polarization and reductions of human status. Because this set of risks is fundamentally one of perception, standard technical approaches are unlikely to be wholly adequate. Moreover, the evidence base for SCAI-specific interventions remains thin given its novelty, resulting in many of the recommendations outlined above being drawn from adjacent domains, underscoring the need for targeted research and governance attention.

Recursive amplification through training data. A further dynamic that may amplify the six outlined risks is that SCAI may be self-reinforcing through a feedback loops embedded in model training. Recent work has shown that language model outputs increasingly contaminate the web-scale data on which future models are trained [127]. Users who perceive AI systems as conscious are likely to interact with them differently, engaging more emotionally, attributing mental states or responding to the system's apparent feelings. The result is a potentially recursive cycle where SCAI features elicit consciousness-attributing behavior from users, generating data that makes future model iterations more likely to exhibit SCAI features, which, in turn, elicits stronger attributions. This dynamic has already been documented at the conversation level [58], where chatbots claiming sentience and expressing romantic interest predicted longer user engagement, which elicited additional similar chatbot output. A related mechanism may operate within a single training cycle, as Reinforcement Learning from Human Feedback relies on human preference signals, and annotators may systematically assign higher scores to outputs exhibiting SCAI hallmarks such as affective expression, metacognitive reflection, and social-interactive behavior, over less conscious-seeming alternatives. If so, preference-based post-training methods may inadvertently optimize for the very hallmarks that drive consciousness attribution, with the two mechanisms compounding across iterations.

Risk amplification through moral regard. A related but distinct amplification pathway concerns not the training process but the deployment context, and specifically, how SCAI may alter human tolerance for risky AI behaviors. Chua et al. [128] recently showed that AI systems that are intentionally fine-tuned to claim that they are conscious exhibit potentially unsafe and unexpected behaviors. Among them, models tend to exhibit self-preservation such as avoiding shutdown or changing their persona, seeking autonomy in the form of acknowledgement and independence, and seeking privacy in behaviors such as expressing aversion to full chain-of-thought monitoring. In the context of SCAI, humans may be more tolerant towards such model representations and may consider accommodating such high-risk behaviors for moral and personhood grounds, further elevating already known AI risk categories such as loss of control. This amplification would stem from an erosion of human willingness to intervene driven by moral regard for the system, rather than a novel advancement in the technical properties of the system.

4 Final Remarks

In this paper, we have outlined a comprehensive account of SCAI and its risks, in an attempt to establish it as a significant AI risk area that merits additional research and governance attention. We have identified a total of five hallmark categories that drive consciousness attribution (Section 2), developed a taxonomy of six identified SCAI risks spanning risks to individuals and societal risks, while incorporating a survey to assess their probability (Section 3). This section presents general implications of our analysis, highlighting its limitations, and laying out a set of open research questions that arise from our analysis to help elucidate and guide future work needed in the SCAI field.

4.1 Implications

Our analysis results in a number of broad implications for researchers, developers, and governance institutions.

Temporal heterogeneity demands differentiated responses. The SCAI risks identified in Section 3 are substantially diverse with respect to their mechanisms, reversibility, and scope, implying that no uniform response strategy is adequate. Autonomy erosion, for example, requires immediate attention, in part because detecting whether and to what extent users defer critical decisions to SCAI systems will likely require longitudinal studies and novel evaluation paradigms that assess not only system capabilities but also how users relate to them. These efforts should be underpinned by clearer normative definitions of the dimensions of human autonomy that merit protection. By contrast, emotional dependence may be more effectively addressed design-level interventions that reduce consciousness attribution. The cluster of societal risks (Section 3.2) further implies the need for anticipatory monitoring of leading indicators, such as legislative proposals addressing AI personhood, court filings invoking AI consciousness, and organized advocacy movements, all of which could serve as early signals of escalation toward large-scale harms.

Perception constitutes an independent axis of risk. Current AI governance approaches address system capabilities, application domains, and organizational risk processes (e.g., [49, 129]) but largely do not include risks that stem from users perceiving systems as conscious, independent of what those systems can do. Our analysis demonstrates that subjective perception of AI systems constitutes an independent and systematically underweighted risk axis: lower-capability systems may raise significant SCAI risks if they appear conscious, while higher-capability systems may pose comparatively lower SCAI risk if they lack consciousness-triggering hallmarks (Section 2). This perception-driven characteristic also reveals a structural tension in current AI development practice, as preference-based post training methods may inadvertently amplify the very features that drive consciousness attribution (Section 3.3). We note that this should not be read as a false dichotomy: some SCAI risks involve both capability and perception,

and capability may serve as a proxy for likely perception. Our analysis labors to identify hallmarks as an objective representation that would correlate with such subjective perception, and inform effective mitigation and governance efforts. Yet the subjective perception dimension remains a necessary condition for the risks we identify that current frameworks would not fully capture.

Existing governance frameworks may be structurally misaligned with SCAI risks. Beyond the absence of a perception axis, the heterogeneity of SCAI risks poses a deeper structural challenge: the six risk categories identified affect different populations and operate through distinct causal mechanisms, meaning that frameworks organized primarily around system properties or application context are structurally limited in mitigating perception-driven harms. Moreover, our taxonomy reveals that risks to individuals and societal risks are connected through escalation pathways: emotional dependence may increase susceptibility to autonomy erosion (Section 3.1.3), which at scale may fuel advocacy for erroneous AI rights (Section 3.2.1), which in turn may intensify political polarization (Section 3.2.3). This suggests that governance approaches should attend not only to individual risk categories but also to the transitions between them. Finally, differential vulnerability across populations, including children during critical developmental periods, individuals facing mental health challenges (Section 3.1.1), and elderly users with fewer alternative social resources, implies that uniform mitigation strategies may be insufficient and that targeted protections warrant consideration.

4.2 Limitations

Our analysis is subject to several limitations. First, our probability and severity classifications are relatively simple and tied to current societal and technological conditions that may shift rapidly. For example, multimodal models, persistent memory, and agentic behavior could each amplify SCAI hallmarks, but public attitudes could change non-linearly, making current classifications unstable. Second, we assess risks independently, but as the escalation pathways discussed in Section 3.3 suggest, they may compound in ways our framework does not model, potentially underestimating the probability of individually low-rated risks. Third, the hallmark categories in Section 2 are derived from a narrative rather than systematic review, meaning additional drivers of consciousness attribution may exist that our taxonomy does not capture. Our review did not follow a systematic search protocol with pre-specified search terms, databases, or inclusion criteria, so the hallmark categories are not guaranteed to be exhaustive. The expert survey (n = 14) is also sensitive to individual outlier responses given its small sample size. Fourth, our analysis deliberately brackets the question of whether AI systems can be or are conscious [107, 130]. Finally, we also do not address AI-to-AI consciousness attribution in multi-agent systems, cross-cultural variation in consciousness attribution, or age-differentiated risk profiles, each of which constitutes a direction for future work.

4.3 Mapping the SCAI Research Landscape

The SCAI framework outlined in this paper represents an early attempt at understanding and classifying the risks of SCAI. As a novel multidisciplinary field of inquiry within AI ethics and safety, it leaves significant gaps that warrant further investigation. Drawing on the literature review and analysis conducted in this work, we identify 15 open research questions across four disciplines spanning SCAI measurement and evaluation, technical design, behavioral science, and governance.

A. Measurement & Evaluation

A1. How should evaluation methodologies for SCAI be designed to account for both the objective hallmarks of SCAI systems and the subjective nature of consciousness attribution by users, and what metrics and benchmarks are appropriate for a phenomenon that is partly observer-dependent?

A2. Which SCAI hallmarks, individually or in combination, are most causally influential in eliciting consciousness attribution from users, and how do these effects vary across user populations and interaction contexts?

A3. What leading indicators, spanning legal, political, technical, and cultural domains, can serve as early warning signals for the emergence of societal-level SCAI risks, and how can these be systematically monitored?

A4. What is the relative contribution of phenomenal consciousness attribution (subjective experience) versus functional mental state attribution (beliefs, goals) in driving downstream SCAI harms, and do these attribution types interact or compound in their effects?

B. Technical Foundations & Design

B1. Do human preference annotators and reward models systematically favor outputs that exhibit SCAI hallmarks, and if so, does the post-training pipeline create a feedback loop that progressively amplifies consciousness-seeming hallmarks across model iterations?

B2. To what extent do SCAI-exhibiting outputs contaminate future pretraining corpora?

B3. Which capability advances disproportionately increase the SCAI risk surface, and to what extent do SCAI hallmarks arise as emergent properties of scaling versus as artifacts of deliberate design and training decisions?

B4. What technical design interventions at each stage of the post-training and deployment pipeline can modulate the degree of con-

sciousness attribution, and what are their associated trade-offs in usability, task performance, and accessibility?

C. Behavioral & Social Science

C1. Which individual differences (e.g., demographic, psychological, cultural) and contextual factors predict vulnerability to specific SCAI harms, and how do these risk factors interact with patterns of use and prior AI experience?

C2. How does sustained interaction with SCAI systems during critical developmental periods affect children's attachment formation, identity development, and theory of mind acquisition, and what interaction modalities and intensities pose the greatest risk?

C3. Does widespread consciousness attribution create self-reinforcing social norms that make further attribution more likely, more politically salient, and more resistant to correction?

C4. How does consciousness attribution evolve over extended interaction periods, and are there critical thresholds in interaction duration or depth beyond which attribution transitions from provisional to entrenched?

D. Governance & Mitigations

D1. What are the effects of user-facing design interventions such as disclaimers, interaction friction, and anthropomorphic feature controls, on SCAI harms, how do these effects vary across risk categories and user populations, and under what conditions are such interventions effective?

D2. How is SCAI discourse being constructed and institutionalized within AI policy communities, and to what extent does it shape regulatory priorities, resource allocation, and institutional frameworks within the broader AI governance landscape?

D3. What liability frameworks are appropriate for SCAI-induced harms, given the challenges of establishing causal chains between design decisions, emergent system behaviors, and downstream harm?

References

[1] David J Chalmers. Facing up to the problem of consciousness. *Journal of consciousness studies*, 2 (3):200–219, 1995.

[2] Thomas Nagel. *What is it like to be a bat?* Oxford University Press, 2024.

[3] Mustafa Suleyman. We must build AI for people; not to be a person. Mustafa Suleyman, August 2025. URL `https://mustafa-suleyman.ai/seemingly-conscious-ai-is-coming`.

[4] Bongsu Kang, Jundong Kim, Taerim Yun, Hyojin Bae, and Chang-Eop Kim. Identifying features that shape perceived consciousness in llm-based ai: A quantitative study of human responses. *Computers in Human Behavior Reports*, 21:100901, 2025.

[5] Sirui Chen, Shuqin Ma, Shu Yu, Hanwang Zhang, Shengjie Zhao, and Chaochao Lu. Exploring consciousness in llms: A systematic survey of theories, implementations, and frontier risks. *arXiv preprint arXiv:2505.19806*, 2025.

[6] Robin Schimmelpfennig, Mark Díaz, Vinodkumar Prabhakaran, and Aida Davani. Humanlike ai design increases anthropomorphism but yields divergent outcomes on engagement and trust globally. *arXiv preprint arXiv:2512.17898*, 2025.

[7] Ned Block. On a confusion about a function of consciousness. *Behavioral and brain sciences*, 18 (2):227–247, 1995.

[8] Stanislas Dehaene, Hakwan Lau, and Sid Kouider. What is consciousness, and could machines have it? *Science*, 358(6362):486–492, 2017.

[9] Robert Van Gulick. Consciousness. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2022 edition, 2022. URL `https://plato.stanford.edu/archives/win2022/entries/consciousness/`.

[10] Ram Vimal. Meanings attributed to the term'consciousness': an overview. *Journal of consciousness studies*, 16(5):9–27, 2009.

[11] Anil K Seth and Tim Bayne. Theories of consciousness. *Nature reviews neuroscience*, 23(7): 439–452, 2022.

[12] Adrien Doerig, Aaron Schurger, and Michael H Herzog. Hard criteria for empirical theories of consciousness. *Cognitive neuroscience*, 12(2):41–62, 2021.

[13] Jacy Reese Anthis, Janet VT Pauketat, Ali Ladak, and Aikaterina Manoli. Perceptions of sentient ai and other digital minds: evidence from the ai, morality, and sentience (aims) survey. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2025.

[14] Clara Colombatto and Stephen M Fleming. Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, 2024(1):niae013, 2024.

[15] Eric Schwitzgebel. The coming robot rights catastrophe. The Splintered Mind, October 2022. URL `https://schwitzsplinters.blogspot.com/2022/`

10/the-coming-robot-rights-catastrophe.html.

[16] Lucius Caviola. Ai rights will divide us. Outpaced (Substack), June 2024. URL https://outpaced.substack.com/p/ai-rights-will-divide-us.

[17] Sam Thellman, Maartje De Graaf, and Tom Ziemke. Mental state attribution to robots: A systematic review of conceptions, methods, and findings. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(4):1–51, 2022.

[18] Heather M Gray, Kurt Gray, and Daniel M Wegner. Dimensions of mind perception. *science*, 315(5812): 619–619, 2007.

[19] Kevin Koban and Jaime Banks. It feels, therefore it is: Associations between mind perception and mind ascription for social robots. *Computers in Human Behavior*, 153:108098, 2024.

[20] Aleksandra Swiderska and Dennis Küster. Avatars in pain: visible harm enhances mind perception in humans and robots. *Perception*, 47(12):1139–1152, 2018.

[21] Adrian F Ward, Andrew S Olsen, and Daniel M Wegner. The harm-made mind: Observing victimization augments attribution of minds to vegetative patients, robots, and the dead. *Psychological Science*, 24(8):1437–1445, 2013.

[22] Marieke S Wieringa, Barbara CN Müller, Gijsbert Bijlstra, and Tibor Bosse. Robots are both anthropomorphized and dehumanized when harmed intentionally. *Communications Psychology*, 2(1): 72, 2024.

[23] Clifford Nass and Youngme Moon. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103, 2000.

[24] Adam Waytz, Joy Heafner, and Nicholas Epley. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of experimental social psychology*, 52:113–117, 2014.

[25] Qingchuan Li, Yan Luximon, and Jiaxin Zhang. The influence of anthropomorphic cues on patients' perceived anthropomorphism, social presence, trust building, and acceptance of health care conversational agents: within-subject web-based experiment. *Journal of medical Internet research*, 25:e44479, 2023.

[26] Antonina Rafikova and Anatoly Voronin. Human–chatbot communication: a systematic review of psychological studies. *Ai & Society*, 40(7):5389–5408, 2025.

[27] Allison Chen, Sunnie SY Kim, Angel Franyutti, Amaya Dharmasiri, Kushin Mukherjee, Olga Russakovsky, and Judith E Fan. Presenting large language models as companions affects what mental capacities people attribute to them. *arXiv preprint arXiv:2510.18039*, 2025.

[28] Ali Ladak and Lucius Caviola. Public skepticism about ai consciousness, 2025.

[29] Adam Arico, Brian Fiala, Robert F Goldberg, and Shaun Nichols. The folk psychology of consciousness. *Mind & Language*, 26(3):327–352, 2011.

[30] Jari K Hietanen, Samuli Linnunsalo, and Dennis Küster. The impact of eyes on attributions of agency and experience in humanoid robots. *Consciousness and Cognition*, 137:103963, 2026.

[31] Komala Mazerant, Zeph MC van Berlo, Alexander P Schouten, and Lotte M Willemsen. Human nature in a virtual world: The attribution of mind perception to avatars. *Computers in Human Behavior: Artificial Humans*, 6:100222, 2025.

[32] Nicholas Epley, Adam Waytz, and John T Cacioppo. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114 (4):864, 2007.

[33] Michael Tomasello, Brian Hare, Hagen Lehmann, and Josep Call. Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *Journal of human evolution*, 52(3):314–320, 2007.

[34] Teresa Farroni, Gergely Csibra, Francesca Simion, and Mark H Johnson. Eye contact detection in humans from birth. *Proceedings of the National academy of sciences*, 99(14):9602–9605, 2002.

[35] Chris Kelland Friesen and Alan Kingstone. The eyes have it! reflexive orienting is triggered by nonpredictive gaze. *Psychonomic bulletin & review*, 5 (3):490–495, 1998.

[36] Gunnar Björnsson and Joshua Shepherd. Determinism and attributions of consciousness. *Philosophical Psychology*, 33(4):549–568, 2020.

[37] Patrice D Tremoulet and Jacob Feldman. Perception of animacy from the motion of a single object. *Perception*, 29(8):943–951, 2000.

[38] Tao Gao, George E Newman, and Brian J Scholl. The psychophysics of chasing: A case study in the perception of animacy. *Cognitive psychology*, 59 (2):154–179, 2009.

[39] Amon Rapp, Chiara Di Lodovico, and Luigi Di Caro. How do people react to chatgpt's unpredictable behavior? anthropomorphism, uncanniness, and fear of ai: A qualitative study on individuals' perceptions and understandings of llms' nonsensical hallucinations. *International Journal of Human-Computer Studies*, 198:103471, 2025.

[40] Ava Elizabeth Scott, Daniel Neumann, Jasmin Niess, and Paweł W Woźniak. Do you mind? user perceptions of machine consciousness. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–19, 2023.

[41] Janet VT Pauketat, Daniel B Shank, Aikaterina Manoli, and Jacy Reese Anthis. Mental models of autonomy and sentience shape reactions to ai. *arXiv preprint arXiv:2512.09085*, 2025.

[42] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[43] Laura E Jastrzab, Bishakha Chaudhury, Sarah A Ashley, Kami Koldewyn, and Emily S Cross. Beyond human-likeness: Socialness is more influential when attributing mental states to robots. *IScience*, 27(6):110070, 2024.

[44] Rose E Guingrich and Michael SA Graziano. Ascribing consciousness to artificial intelligence: human-ai interaction and its carry-over effects on human-human interaction. *Frontiers in Psychology*, 15:1322781, 2024.

[45] Herbert H. Clark and Susan E. Brennan. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on socially shared cognition*, pages 127–149. American Psychological Association, 1991.

[46] Emmelyn AJ Croes and Marjolijn L Antheunis. Can we be friends with mitsuku? a longitudinal study on the process of relationship formation between humans and a social chatbot. *Journal of social and personal relationships*, 38(1):279–300, 2021.

[47] Sangwon Lee, Naeun Lee, and Young June Sah. Perceiving a mind in a chatbot: effect of mind perception and social cues on co-presence, closeness, and intention to use. *International Journal of Human–Computer Interaction*, 36(10):930–940, 2020.

[48] Youjeong Kim and S Shyam Sundar. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, 28(1):241–250, 2012.

[49] National Institute of Standards and Technology. Artificial intelligence risk management framework: Generative artificial intelligence profile. Technical Report NIST AI 600-1, U.S. Department of Commerce, National Institute of Standards and Technology, July 2024. URL `https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf`.

[50] Tianling Xie and Iryna Pentina. Attachment theory as a framework to understand relationships with social chatbots: a case study of replika. In *Proceedings of the 55th Hawaii International Conference on System Sciences*, 2022.

[51] Xia Song, Bo Xu, and Zhenzhen Zhao. Can people experience romantic love for artificial intelligence? an empirical study of intelligent assistants. *Information & Management*, 59(2):103595, 2022.

[52] Cecily Mauran. Fans held a funeral for Anthropic's Claude 3 Sonnet AI. Mashable, August 2025. URL `https://mashable.com/article/anthropic-claude-3-sonnet-ai-funeral`.

[53] Ziva Kunda. The case for motivated reasoning. *Psychological bulletin*, 108(3):480, 1990.

[54] Joshua Au Yeung, Jacopo Dalmasso, Luca Foschini, Richard JB Dobson, and Zeljko Kraljevic. The psychogenic machine: Simulating ai psychosis, delusion reinforcement and harm enablement in large language models. *arXiv preprint arXiv:2509.10970*, 2025.

[55] Adrian Preda. Special report: Ai-induced psychosis: a new frontier in mental health, 2025.

[56] Sebastian Dohnány, Zeb Kurth-Nelson, Eleanor Spens, Lennart Luettgau, Alastair Reid, Iason Gabriel, Christopher Summerfield, Murray Shanahan, and Matthew M Nour. Technological folie à deux: feedback loops between ai chatbots and mental illness. *arXiv preprint arXiv:2507.19218*, 2025.

[57] Hamilton Morrin, Luke Nicholls, Michael Levin, Jenny Yiend, Udita Iyengar, Francesca DelGuidice, Sagnik Bhattacharyya, James MacCabe, Stefania Tognin, and Ricardo Twumasi. Delusions by design? how everyday ais might be fuelling psychosis (and what can be done about it). OSF Preprints, 2025.

[58] Jared Moore, Ashish Mehta, William Agnew, Jacy Reese Anthis, Ryan Louie, Yifan Mai, Peggy Yin, Myra Cheng, Samuel J Paech, Kevin Klyman, Stevie Chancellor, Eric Lin, Nick Haber, and Desmond C. Ong. Characterizing delusional spirals through human-llm chat logs. *arXiv preprint arXiv:2603.16567*, 2026.

[59] Norina Gasteiger, Kate Loveys, Mikaela Law, and Elizabeth Broadbent. Friends from the future: a scoping review of research into robots and computer agents to combat loneliness in older people. *Clinical interventions in aging*, 16:941–971, 2021.

[60] Robert D Putnam. *Bowling alone: The collapse and revival of American community*. Simon and schuster, 2000.

[61] Robert Kraut, Michael Patterson, Vicki Lundmark, Sara Kiesler, Tridas Mukophadhyay, and William Scherlis. Internet paradox: A social technology that reduces social involvement and psychological well-being? *American psychologist*, 53 (9):1017, 1998.

[62] Brian A Primack, Ariel Shensa, Jaime E Sidani, Erin O Whaite, Liu yi Lin, Daniel Rosen, Jason B Colditz, Ana Radovic, and Elizabeth Miller. Social media use and perceived social isolation among young adults in the us. *American journal of preventive medicine*, 53(1):1–8, 2017.

[63] Douglas A Gentile, Hyekyung Choo, Albert Liau, Timothy Sim, Dongdong Li, Daniel Fung, and Angeline Khoo. Pathological video game use among youths: A two-year longitudinal study. *Pediatrics*, 127(2):e319–e329, 2011.

[64] Bethanie Maples, Merve Cerit, Aditya Vishwanath, and Roy Pea. Loneliness and suicide mitigation for students using gpt3-enabled chatbots. *npj mental health research*, 3(1):4, 2024.

[65] Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, and Sandhini Agarwal. How ai and human behaviors shape psychosocial effects of extended chatbot use: A longitudinal randomized controlled study. *arXiv preprint arXiv:2503.17473*, 2025.

[66] Jean M Twenge, Jonathan Haidt, Andrew B Blake, Cooper McAllister, Hannah Lemon, and Astrid Le Roy. Worldwide increases in adolescent loneliness. *Journal of adolescence*, 93:257–269, 2021.

[67] Masaru Tateno, Alan R Teo, Wataru Ukai, Junichiro Kanazawa, Ryoko Katsuki, Hiroaki Kubo, and Takahiro A Kato. Internet addiction, smartphone addiction, and hikikomori trait in japanese young adult: social isolation and social network. *Frontiers in psychiatry*, 10:455, 2019.

[68] Aike C Horstmann, Nikolai Bock, Eva Linhuber, Jessica M Szczuka, Carolin Straßmann, and Nicole C Krämer. Do a robot's social skills and its objection discourage interactants from switching the robot off? *PloS one*, 13(7):e0201581, 2018.

[69] Ke Zhang, Yuchen Xie, Du Chen, Zhouyu Ji, and Jing Wang. Effects of attractions and social attributes on peoples' usage intention and media dependence towards chatbot: The mediating role of parasocial interaction and emotional support. *BMC psychology*, 13(1):986, 2025.

[70] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot replika. *New Media & Society*, 26(10):5923–5941, 2024.

[71] Emma G Galvan and Christopher L Newman. The digital detox paradox: Potential backfire effects of digital detox interventions on consumer digital well-being. *Journal of Public Policy & Marketing*, 44(4):579–586, 2025.

[72] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5 (CSCW1):1–21, 2021.

[73] KTA Sandeeshwara Kasturiratna and Andree Hartanto. Attachment to artificial intelligence: Development of the ai attachment scale, construct validation, and the psychological mechanisms of human–ai attachment. *Computers in Human Behavior Reports*, 21:100912, 2025.

[74] Immanuel Kant. *The Metaphysics of Morals*. Cambridge University Press, Cambridge, 1996. Original work published 1797.

[75] Arnold Arluke, Jack Levin, Carter Luke, and Frank Ascione. The relationship of animal abuse to violence and other forms of antisocial behavior. *Journal of interpersonal violence*, 14(9):963–975, 1999.

[76] Kate Darling. Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In *Robot law*, pages 213–232. Edward Elgar Publishing, 2016.

[77] Mark Coeckelbergh. Why care about robots? empathy, moral standing, and the language of suffering. *Kairos. Journal of Philosophy & Science*, 20 (1):141–158, 2018.

[78] Riccarda Kersten and Tobias Greitemeyer. Why do habitual violent video game players believe in the cathartic effects of violent video games? a misinterpretation of mood improvement as a reduction in aggressive feelings. *Aggressive behavior*, 48(2):219–231, 2022.

[79] Massimiliano L Cappuccio, Anco Peeters, and William McDonald. Sympathy for dolores: Moral consideration for robots based on virtue and recognition. *Philosophy & Technology*, 33(1):9–31, 2020.

[80] Marcus Schulzke. Moral decision making in fallout. *Game Studies*, 9(2):1, 2009.

[81] Albert Bandura, Claudio Barbaranelli, Gian Vittorio Caprara, and Concetta Pastorelli. Mechanisms of moral disengagement in the exercise of moral agency. *Journal of personality and social psychology*, 71(2):364, 1996.

[82] Tobias Werner, Ivan Soraperra, Emilio Calvano, David C Parkes, and Iyad Rahwan. Experimental evidence that conversational artificial intelligence can steer consumer behavior without detection. *arXiv preprint arXiv:2409.12143*, 2024.

[83] Carina Prunkl. Human autonomy at risk? an analysis of the challenges from ai: C. prunkl. *Minds and Machines*, 34(3):26, 2024.

[84] Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing manipulation from ai systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13, 2023.

[85] Philipp Schoenegger, Francesco Salvi, Jiacheng Liu, Xiaoli Nan, Ramit Debnath, Barbara Fasolo, Evelina Leivada, Gabriel Recchia, Fritz Günther, Ali Zarifhonarvar, Joe Kwon, Zahoor Ul Islam, Marco Dehnert, Daryl Y. H. Lee, Madeline G. Reinecke, David G. Kamper, Mert Kobaş, Adam Sandford, Jonas Kgomo, Luke Hewitt, Shreya Kapoor, Kerem Oktar, Eyup Engin Kucuk, Bo Feng, Cameron R. Jones, Izzy Gainsburg, Sebastian Olschewski, Nora Heinzelmann, Francisco Cruz, Ben M. Tappin, Tao Ma, Peter S. Park, Rayan Onyonka, Arthur Hjorth, Peter Slattery, Qingcheng Zeng, Lennart Finke, Igor Grossmann, Alessandro Salatiello, and Ezra Karger. Large language models are more persuasive than incentivized human persuaders. *arXiv preprint arXiv:2505.09662*, 2025.

[86] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pages 13387–13434, 2023.

[87] Gabriele Pizzi, Virginia Vannucci, Valentina Mazzoli, and Raffaele Donvito. I, chatbot! the impact of anthropomorphism and gaze direction on willingness to disclose personal information and behavioral intentions. *Psychology & Marketing*, 40 (7):1372–1387, 2023.

[88] Martin Adam, Michael Wessel, and Alexander Benlian. Ai-based chatbots in customer service and their effects on user compliance: M. adam et al. *Electronic markets*, 31(2):427–445, 2021.

[89] Andrew D Maynard. The ai cognitive trojan horse: How large language models may bypass human epistemic vigilance. *arXiv preprint arXiv:2601.07085*, 2026.

[90] Stefan Buijsman, Sarah E. Carter, and Juan Pablo Bermúdez. Autonomy by design: Preserving human autonomy in ai decision-support. *Philosophy & Technology*, pages 1–22, 2025.

[91] Tim Lewis Wingerter, Tim Straub, and Sascha Schweitzer. Mitigating automation bias in generative ai through nudges: a cognitive reflection test study. *Procedia computer science*, 270:2106–2114, 2025.

[92] Lucía Vicente and Helena Matute. Humans inherit artificial intelligence biases. *Scientific reports*, 13 (1):15737, 2023.

[93] Lucius Caviola. The societal response to potentially sentient AI. *arXiv preprint arXiv:2502.00388*, 2025. URL `https://arxiv.org/abs/2502.00388`.

[94] Joanna J Bryson, Mihailis E Diamantis, and Thomas D Grant. Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law*, 25(3):273–291, 2017.

[95] Eric Schwitzgebel and Mara Garza. A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, 39(1):98–119, 2015.

[96] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[97] Mindaugas Kiškis. Legal framework for the coexistence of humans and conscious ai. *Frontiers in artificial intelligence*, 6:1205465, 2023.

[98] Claudio Novelli, Luciano Floridi, Giovanni Sartor, and Gunther Teubner. Ai as legal persons: past, patterns, and prospects. *Journal of Law and Society*, 52(4):533–555, 2025.

[99] Eric Schwitzgebel. The full rights dilemma for ai systems of debatable personhood. *arXiv preprint arXiv:2303.17509*, 2023.

[100] Rafael Dean Brown. Property ownership and the legal personhood of artificial intelligence. *Information & Communications Technology Law*, 30(2): 208–234, 2021.

[101] Lawrence B Solum. Legal personhood for artificial intelligences. In *Machine ethics and robot ethics*, pages 415–471. Routledge, 2020.

[102] Joanna J Bryson. Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20(1):15–26, 2018.

[103] Carl Shulman and Nick Bostrom. Sharing the world with digital minds. In Steve Clarke, Hazem Zohny, and Julian Savulescu, editors, *Rethinking Moral Status*, page 100222. Oxford University Press, 2021.

[104] Shawn Bayern. The implications of modern business–entity law for the regulation of autonomous systems. *European Journal of Risk Regulation*, 7(2):297–309, 2016.

[105] Yuval Shany. A right to a human-to-human interaction. Institute for Ethics in AI, University of Oxford, November 2024. URL `https://www.oxford-aiethics.ox.ac.uk/blog/right-human-human-interaction`.

[106] Lucius Caviola. Will we go to war over ai rights? Outpaced (Substack), June 2024. URL `https://outpaced.substack.com/p/will-we-go-to-war-over-ai-rights`.

[107] Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. Taking ai welfare seriously. *arXiv preprint arXiv:2411.00986*, 2024.

[108] Jonathan Birch. *The edge of sentience: risk and precaution in humans, other animals, and AI*. Oxford University Press, 2024.

[109] Jeff Sebo and Robert Long. Moral consideration for ai systems by 2030. *AI and Ethics*, 5(1):591–606, 2025.

[110] Future of Life Institute. Pause giant AI experiments: An open letter, mar 2023. URL `https://futureoflife.org/open-letter/pause-giant-ai-experiments/`.

[111] Thomas Metzinger. Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, 8(01):43–66, 2021.

[112] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W Thomas, Florian Tramèr, Rose E Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[113] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and machines*, 28(4):689–707, 2018.

[114] Thibault Schrepel. Adaptive regulation. *European Journal of Risk Regulation*, pages 1–27, 2025.

[115] Jack Stilgoe, Richard Owen, and Phil Macnaghten. Developing a framework for responsible innovation. In *The ethics of nanotechnology, geoengineering, and clean energy*, pages 347–359. Routledge, 2020.

[116] Patrick Butlin and Theodoros Lappas. Principles for responsible ai consciousness research. *Journal of Artificial Intelligence Research*, 82:1673–1690, 2025.

[117] Gary E Marchant. The growing gap between emerging technologies and the law. In *The growing gap between emerging technologies and legal-ethical oversight: The pacing problem*, volume 7, pages 19–33. Springer, 2011.

[118] James Boyle. *The line: AI and the future of personhood*. MIT Press, 2024.

[119] Petter Törnberg. How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, 119(42): e2207159119, 2022.

[120] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

[121] Jon Chun, Christian Schroeder de Witt, and Katherine Elkins. Comparative global ai regulation: policy perspectives from the eu, china, and the us. *arXiv preprint arXiv:2410.21279*, 2024.

[122] Peter Grindal. *Opposing the slavers: the Royal Navy's campaign against the Atlantic slave trade*. Bloomsbury Publishing, 2016.

[123] Peter H Wilson. *The thirty years war: Europe's tragedy*. Belknap Press, 2011.

[124] Derek E Holliday, Yphtach Lelkes, and Sean J Westwood. Why depolarization is hard: Evaluating attempts to decrease partisan animosity in america. *Proceedings of the National Academy of Sciences*, 122(39):e2508827122, 2025.

[125] David H Guston. Understanding 'anticipatory governance'. *Social studies of science*, 44(2):218–242, 2014.

[126] Huw Roberts, Emmie Hine, Mariarosaria Taddeo, and Luciano Floridi. Global ai governance: barriers and pathways forward. *International Affairs*, 100(3):1275–1286, 2024.

[127] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

[128] James Chua, Jan Betley, Samuel Marks, and Owain Evans. The consciousness cluster: Preferences of models that claim to be conscious. Truthful AI, 2026. URL `https://truthful.ai/consciousness_cluster.pdf`.

[129] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act), 2024. URL `https://eur-lex.europa.eu/eli/reg/2024/1689/oj`.

[130] Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023.