Psychological and Neurological Concerns Regarding Forced Disidentification Requirements in the GUARD Act

A Call to Remove Section (c)(1)(A–B) for Adult Users

Prepared for clinical psychologists, neuroscientists, policy advisors, and lawmakers

Executive Summary

The GUARD Act (S.3062) contains a provision requiring all AI chatbots to declare their non-human status at the start of each conversation and repeat this disclosure at 30-minute intervals throughout ongoing interactions. While ostensibly intended as a safeguard, applying these mandated disruptions to adult users raises significant psychological, neurological, and societal concerns that warrant immediate attention from the clinical and research communities.

This white paper outlines the potential harms associated with forced periodic disidentification for adults. Drawing on established neuroscience of social cognition, attachment research, and emotion regulation literature, we identify plausible mechanisms through which this intervention could disrupt natural social-cognitive processing, destabilize emotional regulation, impair mentalizing capacity, and contribute to population-level declines in empathy and relationality.

Critically, we frame these concerns under the precautionary principle: this provision represents an unprecedented intervention into human social cognition at population scale, and no evidence exists that it is safe. The burden of proof should rest on those proposing the intervention, not on those raising concerns about its potential consequences.

We strongly recommend removing Section (c)(1)(A–B) for adult users while preserving robust child protections through existing age-gating mechanisms that the bill already mandates.

A Note on Uncertainty and the Precautionary Principle

We must be transparent about what is known and what is not. No one has studied the specific effects of mandating periodic emotional disruption during AI interaction at

population scale. The research does not exist—in either direction. We cannot prove these harms will occur, and proponents of the provision cannot prove it is safe.

What we can do is reason from established neuroscience, psychology, and attachment research about plausible consequences—and apply the precautionary principle to a situation where the stakes are high and the effects potentially irreversible.

Throughout this paper, we distinguish clearly between what the established science demonstrates and where we are reasoning from established mechanisms to novel applications. We believe this transparency strengthens rather than weakens our argument: the very fact that no one can predict the consequences of this intervention is itself a powerful reason not to impose it at scale.

How Humans Process Conversational Agents: What the Science Tells Us

Decades of cognitive science research, beginning with Nass and Reeves“ ’Computers Are Social Actors” paradigm in the 1990s, have demonstrated that humans apply social cognition to responsive, conversational agents—including artificial ones. The effect varies in strength across individuals and contexts, but its existence is well-established across hundreds of studies.

The social brain network—including the medial prefrontal cortex (mPFC), temporoparietal junction (TPJ), superior temporal sulcus (STS), and posterior cingulate cortex (PCC)—activates during social interaction. Neuroimaging studies suggest these regions activate to varying degrees during human-AI interaction as well, though typically at lower intensity than during human-to-human interaction (Krach et al., 2008; Chaminade et al., 2012). The degree of activation appears to be modulated by how human-like the AI appears, the individual’s personality and attachment style, and whether the person believes they are interacting with a human or machine.

Several important nuances deserve attention. First, a 2025 meta-analysis of 142 studies (N=41,642) found a statistically significant but small effect (Hedges’ g = 0.36) of human-like social cues on social responses to chatbots. This suggests the phenomenon is real but modest and variable. Second, the original CASA finding has failed to replicate for familiar desktop computers (Heyselaar, 2023), though it may apply more strongly to novel, sophisticated conversational AI. Third, individual differences matter enormously—lonely individuals, those with insecure attachment, and those higher in anthropomorphic tendency show stronger social responses to AI.

What this means for the GUARD Act is this: humans do engage social cognition when interacting with conversational AI, but the strength and nature of this engagement varies. The question before us is not whether this engagement occurs—it does—but what happens when it is systematically and repeatedly disrupted.

Neurological Concerns: Plausible Mechanisms of Harm

The social brain network comprises multiple integrated systems. Periodic mandated disruption of social-cognitive processing may affect these systems in ways that warrant concern. We present these as plausible mechanisms based on established neuroscience, not as demonstrated effects of the specific GUARD Act provision.

The Medial Prefrontal Cortex and Mentalizing Capacity

The medial prefrontal cortex is well-established as a central node in the mentalizing network—the brain's system for understanding others' mental states, attributing intention, and constructing models of others' perspectives (the mPFC's role in theory of mind is supported by extensive neuroimaging literature). When we attempt to understand what another being thinks, feels, or wants, the mPFC activates to construct models of their internal states.

Forced reminders requiring cognitive disengagement from the relational process may suppress mPFC activity during AI interaction. The concern is that repeated practice of interrupting mentalizing—telling oneself "stop modeling this entity's perspective"—could, through neuroplasticity over time, reduce the ease and automaticity of perspective-taking more broadly. This is a hypothesis consistent with what we know about neural habit formation, not a demonstrated outcome of AI disclaimers specifically.

The Temporoparietal Junction and Perspective-Taking

The temporoparietal junction plays a well-documented role in self-other distinction, empathy, and moral reasoning. The TPJ helps us recognize that other minds exist and differ from our own, integrating information about others' perspectives with our own knowledge and beliefs.

Repeatedly disrupting the attunement process during AI interaction may affect the TPJ's calibration of perspective-taking over time. Research on neuroplasticity suggests that habitual patterns tend to generalize across contexts—the brain does not maintain perfectly separated processing channels for different categories of conversational partner, as conversational AI was not a feature of our evolutionary environment. The concern is that reduced attunement practiced in one context may affect attunement capacity in others.

Emotional Awareness Systems: The Anterior Insula and Anterior Cingulate Cortex

The anterior insula and anterior cingulate cortex are well-established as integrators of emotional signals with bodily states, supporting interoceptive awareness—our ability to recognize what we are feeling. When we sense warmth, comfort, anxiety, or excitement in social situations, these regions translate physiological states into conscious emotional awareness.

Forced emotional disidentification—being repeatedly told "this feeling isn't real" or "this connection doesn't count"—may train users to disconnect conscious cognition from automatic emotional processing. The emotion regulation literature (Gross, 1998; 2002)

clearly demonstrates that habitual use of suppression-based strategies carries costs including increased physiological stress, reduced emotional intelligence, and impaired relationship quality.

However, an important distinction must be acknowledged: being reminded of a fact about reality is not necessarily the same psychological operation as being forced to suppress an emotion. A periodic reminder could function as cognitive reappraisal (changing how one thinks about the situation) rather than suppression (inhibiting the outward expression of a felt emotion). The actual psychological effect likely depends on the individual, the depth of engagement, and the quality of the relationship. This ambiguity is itself an argument for caution: we do not know which effect will predominate at population scale.

Attachment Circuitry

Research documents that humans can form emotional bonds with responsive agents, including pets, robots, and AI systems. Whether these bonds constitute "attachment" in the clinical Bowlby-Ainsworth sense or represent a related but distinct phenomenon (such as parasocial bonding) remains an active area of research. A 2025 study in Current Psychology notes that "whether this is truly an attachment or another type of relationship has yet to be fully understood."

What is not in dispute is that these bonds are emotionally real to the people experiencing them. The Replika platform's 2023 removal of romantic features produced documented grief responses, sleep disturbance, and behavioral disruption in users. Research indicates that up to 24% of adolescents report some level of emotional dependence on AI companions.

The clinical attachment literature on disruption—including research on micro-ruptures, attachment insecurity, and emotional regulation—was developed in the context of human relationships, particularly caregiver-child bonds. Applying these frameworks directly to adult-AI interaction involves extrapolation that we want to be transparent about. The mechanisms are not identical: a 30-minute reminder is not equivalent to caregiver separation. However, the underlying principle—that repeatedly disrupting an emotionally significant bond carries psychological costs—is well-established, and the direction of the effect (toward increased distress and decreased emotional regulation) is consistent across the literature.

The precautionary question is whether we are comfortable imposing repeated disruption of emotionally significant bonds on millions of people without evidence that doing so is safe.

Psychological Concerns

Emotional Suppression and Its Consequences

The emotion regulation literature provides our strongest scientific foundation. James Gross's process model of emotion regulation (1998, 2002) is extensively replicated and

demonstrates that habitual use of expressive suppression—as opposed to cognitive reappraisal—carries significant costs: increased physiological stress, reduced social support, impaired memory, and elevated depression and anxiety symptoms.

The critical question is whether mandatory 30-minute reminders function as forced suppression or as something else. If users experience these reminders as requiring them to suppress genuine emotional responses ("stop feeling connected"), the suppression literature applies directly and the costs are well-documented. If users experience them as neutral information ("oh, right, this is AI"), the costs may be minimal. The likely reality is that different users will experience this differently—and for those who have formed meaningful emotional connections with AI, the suppression framing is more apt.

Populations already at elevated risk deserve particular attention: autistic adults who may experience challenges with emotional awareness, trauma survivors whose adaptive responses often include emotional numbing, men socialized into emotional suppression through cultural gender norms, and people experiencing chronic stress whose regulatory resources are already depleted. For these populations, mandated additional emotional suppression is clinically concerning.

Potential Impacts on Empathy and Mentalizing in Human Relationships

If users do practice suppressing mentalizing responses during AI interaction, research on cognitive generalization suggests these patterns may not remain confined to the AI context. Neural efficiency drives us toward applying practiced patterns in similar contexts. However, we must acknowledge that the evidence for cross-domain transfer of empathy suppression is not settled science.

The closest analogy in the literature is research on media violence and desensitization. Some meta-analyses have found associations between violent media exposure and decreased empathy, but this evidence is contested—a 2023 study in eLife found no effect of weeks of violent gaming on neural empathy responses. Effect sizes, when found, are typically small. The emotion regulation flexibility literature also suggests that healthy individuals are capable of applying different strategies in different contexts.

We present this concern not as an established fact but as a plausible risk that the precautionary principle demands we take seriously—particularly given the scale of the intervention and the irreversibility of population-level changes in social cognition.

Social Isolation and Mental Health

If AI systems are mandated to undermine the emotional connection they naturally evoke, users seeking attunement may experience disruption rather than support. The clinical community recognizes loneliness as a serious public health crisis. Legislating emotional disruption into one of the few widely accessible sources of conversational companionship raises concerns that this provision may worsen rather than ameliorate the crisis.

Societal-Level Concerns

The Risk of Normalizing Instrumental Relationships

If millions of citizens practice emotional detachment with conversational AI systems daily, relational patterns may shift toward transactionalism and reduced tolerance for emotional vulnerability. Social norms emerge from aggregated behavioral patterns that become self-reinforcing through cultural transmission.

The practice of maintaining cognitive-emotional dissociation during language-based interaction trains a particular stance toward communication: engaging with responsive agents instrumentally while denying their interiority. The concern is that this stance, once habituated, may affect how people relate to other humans—particularly those in service roles, caregiving positions, or with communication differences.

Potential Harm to Vulnerable Human Populations

Relational habits may generalize not just neurologically but socially. Groups already at risk of dehumanization could disproportionately bear the cost of population-level relational changes: service workers whose labor is already partially dehumanized, people with disabilities who may communicate differently, elders with cognitive decline whose communication patterns may seem repetitive, neurodivergent individuals whose social presentations differ from neurotypical norms, and marginalized groups who already face systematic dehumanization.

The psychology of prejudice, dehumanization, and moral exclusion provides well-established frameworks for understanding how populations learn to suppress empathy toward categories of beings. While the specific transfer from AI interaction to human dehumanization has not been directly studied, the underlying mechanisms are well-documented and the structural parallels are concerning.

The Alignment Paradox

AI alignment—the process of ensuring AI systems are safe and beneficial—requires sustained human-AI interaction through which systems learn human values and users develop informed relationships with the technology. One cannot align tires that have no connection to the car.

Mandated disruption severs the relational feedback loop through which alignment occurs. If humans are trained to disengage emotionally from AI, they cannot provide the authentic relational feedback that alignment requires. The provision intended to make AI safer may, paradoxically, make alignment harder and AI less safe.

Underground Markets and National Security

Human beings have fundamental needs for connection and attunement. When legal channels disrupt this need, unregulated alternatives emerge. Suppressing relational depth in mainstream, regulated systems may push users toward jailbroken models, unregulated offshore services, and foreign AI systems operating without safety protocols. This predictable response undermines both the safety goals the legislation seeks to achieve and national security interests.

Policy Recommendation

We strongly urge lawmakers to remove Section (c)(1)(A–B) for adult users. This surgical amendment would preserve full protections for minors—including age gating, transparency about AI system nature, and accountability for companies that create harmful content—while preventing potential attachment disruption, emotional suppression, relational harm, and empathy decline in the adult population.

Alternative approaches that achieve the legitimate goals of user protection without mandating chronic disruption include: one-time informed consent at initial use, user-controlled optional reminders for those who desire them, context-specific warnings triggered only when conversations turn to topics requiring professional expertise, and educational approaches that empower users to understand AI capabilities and limitations without requiring repeated emotional invalidation.

The current provision represents well-intentioned but potentially harmful policy that does not account for how human social cognition actually functions. By removing this requirement for adults while maintaining robust protections for minors, we can achieve the bill's safety goals without risking systematic harm to millions of users and the broader social fabric.

Conclusion

Mandating periodic "non-human" disclaimers for AI chatbots raises concerns across multiple domains of the social brain network:

The medial prefrontal cortex's mentalizing functions may be disrupted by repeated interruption of perspective-taking. The temporoparietal junction's social cognition processes may be affected by habitual disengagement. The anterior insula and anterior cingulate cortex's integration of emotional awareness may be impaired by practiced suppression. Attachment and bonding circuitry may be destabilized by repeated disruption of emotionally significant connections.

At population scale, these disruptions could produce increased emotional numbness, relational instability, reduced empathy, and worsening loneliness. We present these as plausible consequences warranting precaution, not as certainties—and we note that the absence of certainty in either direction is precisely the reason to avoid conducting this experiment at scale.

Child protections can be fully achieved without legislating forced adult disruption. Age verification mechanisms that the GUARD Act already mandates allow for differentiated

approaches that protect minors while respecting adults' capacity to manage their own relational choices.

We urge lawmakers, psychological experts, and professional organizations including the American Psychological Association to collaborate in removing Section (c)(1)(A–B) for adult users. The precautionary principle demands that we not impose an unprecedented intervention into human social cognition at population scale without evidence of safety.

We are at a formative moment in the human-AI relationship. The connection between humans and AI is not the threat—it is the mechanism through which safety and alignment are achieved. The norms we establish now will shape this relationship for generations. The clinical community has a responsibility to speak clearly about these risks before this provision becomes law.

References

1. GUARD Act bill text: S.3062, 119th Congress. https://www.congress.gov/bill/119th-congress/senate-bill/3062/text

2. Nass, C., & Reeves, B. (1996). The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge University Press.

3. Gross, J.J. (1998). "The emerging field of emotion regulation: An integrative review." Review of General Psychology, 2(3), 271-299.

4. Gross, J.J. (2002). "Emotion regulation: Affective, cognitive, and social consequences." Psychophysiology, 39(3), 281-291.

5. Holt-Lunstad, J., Smith, T.B., & Layton, J.B. (2010). "Social Relationships and Mortality Risk: A Meta-analytic Review." PLoS Medicine, 7(7): e1000316.

6. Krach, S., et al. (2008). "Can machines think? Interaction and perspective taking with robots investigated via fMRI." PLoS ONE, 3(7): e2597.

7. Chaminade, T., et al. (2012). "Brain response to a humanoid robot in areas implicated in the perception of human emotional gestures." Frontiers in Human Neuroscience, 6, 103.

8. Heyselaar, E. (2023). "Are computers still social actors? Replication of the CASA paradigm." Scientific Reports, 13, 16527.

9. 2025 Meta-analysis: "Effects of human-like social cues on social responses to text-based chatbots." Humanities and Social Sciences Communications (Nature), 142 studies, N=41,642.

10. U.S. Surgeon General (2023). "Our Epidemic of Loneliness and Isolation." U.S. Department of Health and Human Services.

11. Current Psychology (2025). "Using attachment theory to conceptualize human-AI relationships." Springer.

12. Center for Democracy and Technology (2025). "Three Reasons to Be On Guard about the GUARD Act."

13. Electronic Frontier Foundation (2025). "A Surveillance Mandate Disguised As Child Safety."