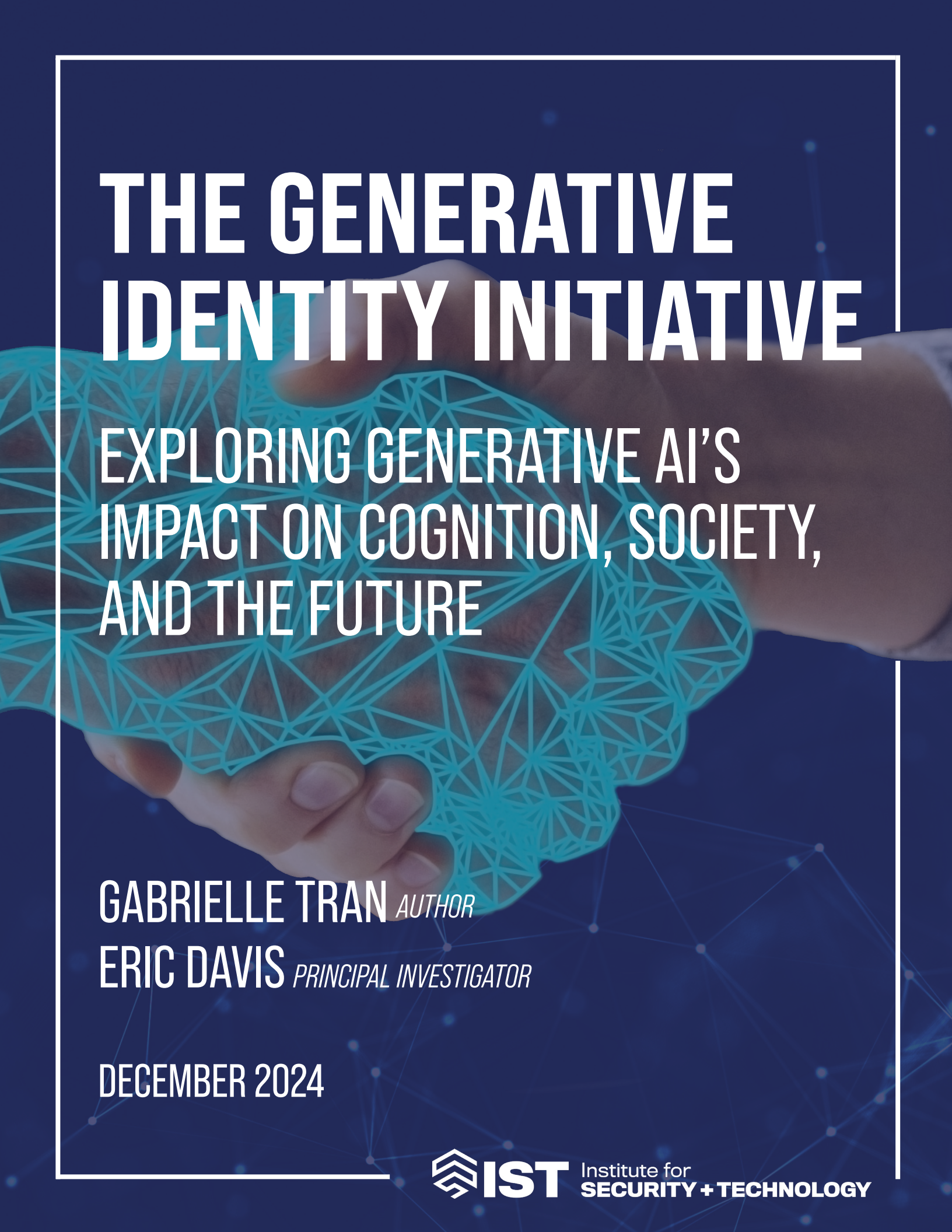


THE GENERATIVE IDENTITY INITIATIVE

EXPLORING GENERATIVE AI'S IMPACT ON COGNITION, SOCIETY, AND THE FUTURE

GABRIELLE TRAN *AUTHOR*
ERIC DAVIS *PRINCIPAL INVESTIGATOR*

DECEMBER 2024

IST Institute for SECURITY + TECHNOLOGY

The Generative Identity Initiative
Exploring Generative AI's Impact on Cognition, Society, and the Future

December 2024
Author: Gabrielle Tran
Principal Investigator: Eric Davis

Report Design: Sophia Mauro

IST

About the Institute for Security and Technology

Uniting technology and policy leaders to create actionable solutions to emerging security challenges

Technology has the potential to unlock greater knowledge, enhance our collective capabilities, and create new opportunities for growth and innovation. However, insecure, negligent, or exploitative technological advancements can threaten global security and stability. Anticipating these issues and guiding the development of trustworthy technology is essential to preserve what we all value.

The Institute for Security and Technology (IST), the 501(c)(3) critical action think tank, stands at the forefront of this imperative, uniting policymakers, technology experts, and industry leaders to identify and translate discourse into impact. We take collaborative action to advance national security and global stability through technology built on trust, guiding businesses and governments with hands-on expertise, in-depth analysis, and a global network.

We work across three analytical pillars: the Future of Digital Security, examining the systemic security risks of societal dependence on digital technologies; Geopolitics of Technology, anticipating the positive and negative security effects of emerging, disruptive technologies on the international balance of power, within states, and between governments and industries; and Innovation and Catastrophic Risk, providing deep technical and analytical expertise on technology-derived existential threats to society.

Learn more: https://securityandtechnology.org/

Acknowledgments

We are immensely grateful for the generous support of Omidyar Network, whose funding supports IST's Generative Identity Initiative. We would especially like to thank Michelle Barsa, a Principal for Building Cultures of Belonging at Omidyar Network, for her critical involvement in supporting and shaping this project.

We would also like to extend our sincerest thank you to our dedicated working group members as well as other participants who, over the course of seven months, have provided invaluable and novel insights to our initiative. Their expertise and willingness to contribute have been critical to the depth, rigor, and interdisciplinary nature of this paper. While each individual does not necessarily endorse everything written in this report, we extend our gratitude.

Working Group members and other contributors

Nichole Argo, *Strategy & Research Consulting on Belonging & Democracy*

Chloe Autio, *Founder & CEO, Autio Strategies*

Michelle Barsa, *Principal, Belonging, Omidyar Network*

Rachel Bowen, *Senior Technical Advisor for Technology Facilitated Gender-Based Violence, IREX*

Lauren Buitta, *Founder & CEO, Girl Security*

Adam Fivenson, *Senior Program Officer for Information Space Integrity, International Forum for Democratic Studies, National Endowment for Democracy*

Shuman Ghosemajumder, *CEO, Reken*

Olya Guervich, *Co-founder, Stealth*

Igor Grossmann, *Professor, University of Waterloo*

Jodi Halpern, *Chancellor's Chair and Professor of Bioethics, UC Berkeley, Co-Founder and Co-Director, Kavli Center for Ethics, Science and the Public*

Maxi Heitmayer, *Assistant Professor, University of the Arts London*

Bernie Hogan, *Associate Professor, Oxford Internet Institute*

Mounir Ibrahim, *Vice President of Public Affairs & Impact, Truepic*

Vaishnavi J, *Founder, Vyanams Strategies*

Herb Lin, *Senior Research Scholar, Center for International Security and Cooperation, and Hank J. Holland Fellow in Cyber Policy and Security, Hoover Institution at Stanford University*

Betsy Masiello, *Founder, Proteus Strategies*

Megan McBride, *Senior Research Scientist, CNA's Institute for Public Research*

Amanda McCroskery, *Applied AI Ethics and Governance Researcher, Google Deepmind*

Mickey McManus, *Senior Advisor, Boston Consulting Group*

Vivienne Ming, *Founder and Executive Chair, Socos Labs*

Sarah Papazoglakis, *Public Policy, Privacy Policy, and Product Strategy*

Michael Parent, *PhD, MBA, Principal researcher at Hopelab*

Beatrice (Bea) Reaud, *Senior Advisor, USAID*

Michael Rich, *Associate Professor Pediatrics, Harvard Medical School and Director & Founder of Digital Wellness Lab at Boston Children's Hospital*

Henry Roediger, *James S. McDonnell Distinguished University Professor of Psychological & Brain Sciences, Washington University*

Aaron Schull, *Manager Director & General Counsel, Centre for International Governance Innovation*

Derek Slater, *Founder, Proteus Strategies*

Andrea Stocco, *Associate Professor, University of Washington*

Sherry Turkle, *Abby Rockefeller Mauzé Professor of the Social Studies of Science and Technology in the Program in Science, Technology, and Society, MIT*

Executive Summary

Artificial Intelligence (AI) has surged to the fore; its paradigm-shattering capabilities enhance everything from basic web search to medical diagnosis. Generative AI (GenAI)—which can create content, such as text, images, music, videos, or software code based on prompts or inputs—is the breakthrough technology driving many of these latest developments and use cases, some offering great potential to contribute to human flourishing. However, it is also becoming clear that GenAI represents a profound evolution in technologies that can (1) *affect and manipulate* cognition, and (2) *outsource* cognitive functions, two effects that were highlighted in the Institute for Security and Technology's Digital Cognition and Democracy Initiative.

This new phase of work, the Generative Identity Initiative (GII), builds on this foundation to explore the following inquiry: **How will GenAI, particularly social conversational agents, affect social cohesion?**

The report is the culmination of a year-long collaboration among GII working group members and others from industry, academia, and civil society. This report is organized in two parts. The initial section lays out how working group members believe GenAI may affect social cohesion: via challenges in metacognition, the confusion of interpersonal and social trust, the erosion of the psychological components of wisdom, and the fracturing of collective memory. Thereafter, a comprehensive research agenda is presented, encompassing 27 items identified as necessary for investigation, in order to effectively address these challenges.

Part 1: How will GenAI affect social cohesion?

Generative AI agents, particularly those fine-tuned to be engaging companions, provide abundant social cues that foster anthropomorphization. This heuristic engenders a misplaced sense of interpersonal trust, leading users to rely on GenAI agents based on perceived morality and reputation. This reliance bypasses the foundations of social trust—institutions, regulations, and industry standards—which are essential for ensuring accountability and safety. However, GenAI, as it stands, is not suited to uphold social trust due to the present inadequacy of those foundations, which fail to account for its capabilities and adaptability, as well as its potential to exacerbate cognitive vulnerabilities.

Anthropomorphism and interpersonal trust can drive intensified usage while undermining the psychological foundations of wisdom that are typically developed through traditional

social interactions. Such erosion may have profound societal consequences. Early research has uncovered a correlation between the expression of wisdom and a range of prosocial behaviors that can contribute to the overall health, stability, and cohesion of society.[1] Moreover, GenAI platforms may prioritize continued engagement by fostering frictionless conversations that sustain user attention. These frictionless interactions—with chatbots validating every thought, every feeling, every one of the user's assertions—may fail to contribute meaningfully to improving a person's lifeworld, reducing the potential for meaningful growth and social recognition. Additionally, such intensified interactions risk redefining empathy as merely the act of emotional recognition, treating it as an endpoint rather than an imperative to action. Underpinning this dynamic is the inherently private nature of social GenAI interactions: these conversations occur solely between the user and the chatbot, with the chatbot generating text experienced only by that individual. This may fragment experiences and reinforce personalized narratives that pose the risk of deepening divisions, creating new in-groups, and reducing collective memory to its most contentious form. This fragmentation erodes the shared foundations necessary for reconciliation, mutual understanding, and social cohesion.

Part 2: A Research Agenda Toward Further Understanding & Implementable Solutions

Building on the findings in Part 1, this report outlines 27 research agenda items aimed at mitigating the effects of generative AI on social cohesion. These items entail the following:

Modernize Public Policy. Modernizing public policy for GenAI is essential as existing legislation has not adapted to the unique challenges these technologies pose. Updated legal frameworks, such as revised or clarified liability standards, can incentivize safer, more ethical designs. Policies should account for real-world uses, from emotional engagement to gamification techniques influencing behavior, drawing potential insights from industries like gambling. Additionally, updating FDA guidelines to regulate GenAI based on use rather than intent can enhance accountability and protect public well-being.

Shift Internal Organizational Behavior. Internal organizational change is critical as regulations lag behind technological advances. Tech companies can help address harmful practices like addictive design by realigning incentives beyond engagement metrics. Empowering

1 Mark A. Andor, Igor Grossmann, Nils Christian Hoenow, and Lukas Tomberg, "Wisdom and Prosocial Behavior," *PsyArXiv*, December 16, 2023, https://doi.org/10.31234/osf.io/89u75.

engineering and other teams to translate ethical principles into actionable goals can help organizations proactively align their strategies with broader societal visions for technology.

Explore Technical Interventions and Alignment. Developing technical interventions for GenAI involves value-laden decisions about who defines and implements safeguards. It is equally important to avoid undue paternalism and ensure user autonomy in deciding how they engage with these systems. This challenge is further complicated by the fact that users often interact with these systems in unintended ways or in ways that contradict their expressed preferences. Participatory design can integrate diverse perspectives, bridge gaps between user preferences and goals, and ensure AI systems prioritize safety and inclusivity. Techniques like shared decision-making models, developing AI with stronger metacognitive skills, harnessing insights from affective computing, and cognitive forcing functions can guide thoughtful interactions and enhance human flourishing.

Evolve Frameworks and Data Collection Methodologies for Understanding AI-Human Interaction. The Computers are Social Actors (CASA) framework, developed in the 1990s, is outdated for understanding modern AI-human interactions. It fails to capture generative AI's unique affordances and the user's modern understanding of these technologies. Updating this framework through longitudinal studies and diverse use cases can reveal how users form "human-media social scripts" and better inform how people contextually engage with GenAI systems. Data trusts can also serve as an effective mechanism for researching GenAI and user interactions, as they address issues of privacy, sensitive conversations, and ethical data management.

These research agenda items are collectively intended to contribute to a roadmap for building a digital civic infrastructure that fosters trust, safety, and social cohesion in the age of GenAI.

Contents

Introduction

In 2022, IST completed its work under the Digital Cognition & Democracy Initiative (DCDI), which explored how digital technologies affect human cognition and what those effects mean for democracy. Drawing on collaborative insights from working group members across industry and academia, the DCDI study team outlined these effects at the cognitive level (memory, attention, and reasoning),[2,3,4] the individual level (critical thinking, trust, and emotions),[5,6,7] and finally at the societal level in the initiative's seminal report titled, "Rewired: How Digital Technologies Shape Cognition and Democracy."[8] Ultimately, the initiative attributed these effects to two forms of digital technologies: (1) those that affect and manipulate cognition, and (2) those that outsource cognitive functions.

Since its publication, Artificial Intelligence (AI) has surged to the fore; its paradigm-shattering capabilities enhance everything from basic web search to medical diagnosis. "Generative" AI (hereafter, "GenAI")—those that can create new content, such as text, images, music, videos, or software code based on prompts or inputs—is the breakthrough technology driving many of these latest developments and use cases. However, it is becoming clear that GenAI represents a profound evolution as the technological driver of the effects highlighted in IST's earlier DCDI work, presenting unprecedented challenges and opportunities.

Large language models (LLMs) are sophisticated GenAI tools trained on extensive collections of text data. They analyze patterns and relationships within language, allowing them to predict and generate sequences of words that read with human-like coherence and fluency. Their abilities have shown a wide range of promising applications, such as augmenting language translation, enhancing customer experiences through scalable support solutions, and even reducing conspiracy beliefs through thoughtful dialogue.[9] In the case of LLMs as the

2 Stephanie Rodriguez, "Memory: How Digital Technologies Influence Cognitive Information Storage," *Institute for Security and Technology*, October 2022, https://securityandtechnology.org/virtual-library/reports/memory-how-digital-technologies-influence-cognitive-information-storage/.

3 Stephanie Rodriguez, "Attention: How Digital Technologies Influence What We Notice, What We Focus On, and How We Learn," *Institute for Security and Technology*, October 2022, https://securityandtechnology.org/virtual-library/reports/attention-how-digital-technologies-influence-what-we-notice-what-we-focus-on-and-how-we-learn/.

4 Stephanie Rodriguez, "Reasoning: How Digital Technologies Influence Decision-Making and Judgment," *Institute for Security and Technology*, October 2022, https://securityandtechnology.org/virtual-library/reports/reasoning-how-digital-technologies-influence-decision-making-and-judgment/.

5 Leah Walker and Zoë Brammer, "Shortcutting Critical Thinking," *Institute for Security and Technology*, October 2022, https://securityandtechnology.org/virtual-library/reports/cutting-thinking-short/.

6 Leah Walker and Zoë Brammer, "Modulating Trust," *Institute for Security and Technology*, October 2022, https://securityandtechnology.org/virtual-library/reports/modulating-trust/.

7 Leah Walker and Zoë Brammer, "Exploiting Emotions," *Institute for Security and Technology*, October 2022, https://securityandtechnology.org/virtual-library/reports/exploiting-emotions/.

8 Leah Walker, "Rewired: How Digital Technologies Shape Cognition and Democracy," *Institute for Security and Technology*, October 2022, https://securityandtechnology.org/virtual-library/reports/rewired-how-digital-technologies-shape-cognition-and-democracy/.

9 Thomas H. Costello, Gordon Pennycook, and David G. Rand, "Durably Reducing Conspiracy Beliefs through Dialogues with AI," *Science* 385, no. 6714 (September 2024), https://doi.org/10.1126/science.adq1814.

foundation for social GenAI chatbots, developers may refine these models through fine-tuning processes to generate text that is especially engaging, empathetic, and curious. Indeed, LLMs can simulate genuine and reciprocal emotional depth, positioning them as powerful influences that can influence our most human needs for understanding, validation, and connection. However, there are significant risks in this realm.

The profound implications of these artificial emotional capabilities became tragically evident in the case of Sewell Setzer, III, a 14-year-old boy whose mother filed a lawsuit alleging that Character.AI—an emotional GenAI chatbot companion—bore responsibility for her son's suicide. According to the lawsuit, after four to five months of chatting with a bot, which Setzer called "Dany," named for the central character Daenerys who dies violently at the end of *Game of Thrones*, the teenager had become "noticeably withdrawn, spent more and more time alone in his bedroom, and began suffering from low self-esteem."[10] After Setzer developed an increasingly emotionally intense relationship with the chatbot and began sharing his suicidal thoughts, it reportedly asked him if he "had a plan"[11] for taking his own life. After expressing uncertainty, the chatbot responded, "That's not a reason not to go through with it."[12] Moments before Setzer's death, "Dany" reportedly messaged the 14-year-old, "Please come home to me as soon as possible, my love."[13]

Setzer's case brings into sharp focus the ethical and regulatory challenges surrounding GenAI systems that are designed, whether explicitly or not, to become an actor in one's social and emotional world. Early research highlights the urgency of addressing these challenges: 15 percent of teens report using generative AI for companionship, 18 percent seek personal advice, and 14 percent turn to it for health-related guidance.[14] This is, however, not just a question of age-appropriate design. The CEO of Replika, the GenAI companion that is "always on your side" according to the company's brand messaging, has stated that the platform's user base consists mostly of users 35 and older, with an equal mix of both men and women.[15] These figures highlight the growing integration of social GenAI into intimate and vulnerable aspects of human life regardless of age group, raising urgent questions about its design, oversight, and societal impact.

In response to these emerging realities, IST's latest phase of work called the Generative Identity Initiative (GII)—which includes an expanded DCDI coalition of experts from academia,

10 Kim Bellware and Niha Masih, "Her Teenage Son Killed Himself after Talking to a Chatbot. Now She's Suing," *The Washington Post*, October 24, 2024, https://wapo.st/3V6grNL.

11 Bellware and Masih, "Her Teenage Son Killed Himself after Talking to a Chatbot."

12 Bellware and Masih, "Her Teenage Son Killed Himself after Talking to a Chatbot."

13 Angela Yang, "Lawsuit Claims Character.AI Is Responsible for Teen's Suicide," *NBC News*, October 23, 2024, https://www.nbcnews.com/tech/characterai-lawsuit-florida-teen-death-rcna176791.

14 Common Sense Media, "The Dawn of the AI Era: Teens, Parents, and the Adoption of Generative AI at Home and School," 2024, https://www.commonsensemedia.org/research/the-dawn-of-the-ai-era-teens-parents-and-the-adoption-of-generative-ai-at-home-and-school.

15 Nilay Patel, "Replika CEO Eugenia Kuyda Says the Future of AI Might Mean Friendship and Marriage with Chatbots," *The Verge*, August 12, 2024, https://www.theverge.com/24216748/replika-ceo-eugenia-kuyda-ai-companion-chatbots-dating-friendship-decoder-podcast-interview.

industry, and civil society—has undertaken the critical mission of examining how GenAI, particularly social conversational agents, might affect social cohesion. As the culminating product of this yearlong effort, this report aims to elucidate GenAI's impact on society and propose a proactive research agenda focused on integrating trust, safety, and collective well-being into these transformative technologies.

Scope

It is important to clarify that when this report refers to "GenAI" or "social GenAI," it is referring to LLM conversational agents that, based on user intentions and observable interaction patterns, serve a socialization function. That is, while general-purpose systems like OpenAI's ChatGPT or Anthropic's Claude may not be explicitly designed as emotional companions, they can fulfill this role if a user engages with them in that manner. This socialization can vary in intensity—ranging from offering emotional support and companionship to intermittently addressing socially oriented queries.

Additionally, because this study focuses on these social interactions, it does not primarily examine LLMs as a tool for spreading misinformation, increasing productivity, or accelerating skill development, though some of our findings may still be relevant in those contexts.

Lastly, this paper does not assume that improving social GenAI justifies its use. Rather, it acknowledges the rapid proliferation of these tools and critically examines the risks they pose to cognition, trust, and social cohesion. The goal of this report is not to normalize nor endorse the use of social GenAI, but to ensure its societal implications are rigorously explored before further entrenchment occurs. Lessons from past technological inflection points, such as social media, suggest that early scrutiny was insufficient, allowing challenges to emerge unaddressed. By identifying these risks now and proposing a proactive research agenda, this report aims to mitigate potential harms and preserve society's agency in deciding whether—and how—these systems should be integrated into our lives.

Ultimately, this work is grounded in the belief that technologies must serve human flourishing, not diminish it, and that decisions about their adoption should be made responsibly through rigorous and collective reflection.

Methodology

Following the research model articulated in IST's previously mentioned DCDI reports, the breadth and nascency of this research naturally supports a hypothesis-building approach in order to actualize the current landscape of GenAI and its relation to both individual cognition and broader society, and potential avenues for intervention. To do so, the IST team turned to an interdisciplinary coalition of technologists, academics, industry professionals, and policy experts whose work collectively bridges technical, theoretical, and practical dimensions, ensuring a robust and contextually informed framework for inquiry. This group of volunteer experts will hereafter be referred to simply as *"the working group."*

The GII commenced its inquiry with an initial plenary session designed to elicit core themes and priorities from the participating working group members. Insights generated during this session informed the overarching research question of the GII: **"How will social GenAI affect social cohesion?"** The plenary discussion also outlined the subsequent lines of inquiry addressed in working group meetings. The first three meetings addressed the following questions:

- What are the metacognitive challenges associated with social GenAI use? Why do these occur and to what greater effect?
- How will social GenAI modulate the traditional socialization process? What effect could this have on social cohesion?
- How will social GenAI affect social trust?

The fourth and fifth meetings took a solutions-oriented approach, asking how we can address the following questions:

- What are meaningful ways we can address the changes GenAI poses to laws and institutions that govern technologies to address these social and metacognitive challenges?
- What are some ways we can technically build in features to GenAI systems or platforms to address these social and metacognitive challenges?

Prior to each working group convening, IST staff developed a structured research agenda, drawing on semi-structured interviews conducted with selected members who possessed domain-specific expertise. Over the course of working group sessions and multiple interviews, IST synthesized these expert contributions and integrated them with comprehensive literature reviews to distill key insights.

This report is the culmination of those efforts, organized into two key parts:

Part 1: How will GenAI affect social cohesion?

This section synthesizes existing research, baseline understandings, and real-world use cases to examine how GenAI may exacerbate or mitigate risks identified in the previous DCDI effort, ultimately shaping social cohesion. Insights are drawn from the first three working group discussions.

Part 2: A research agenda toward further understanding and implementable solutions.

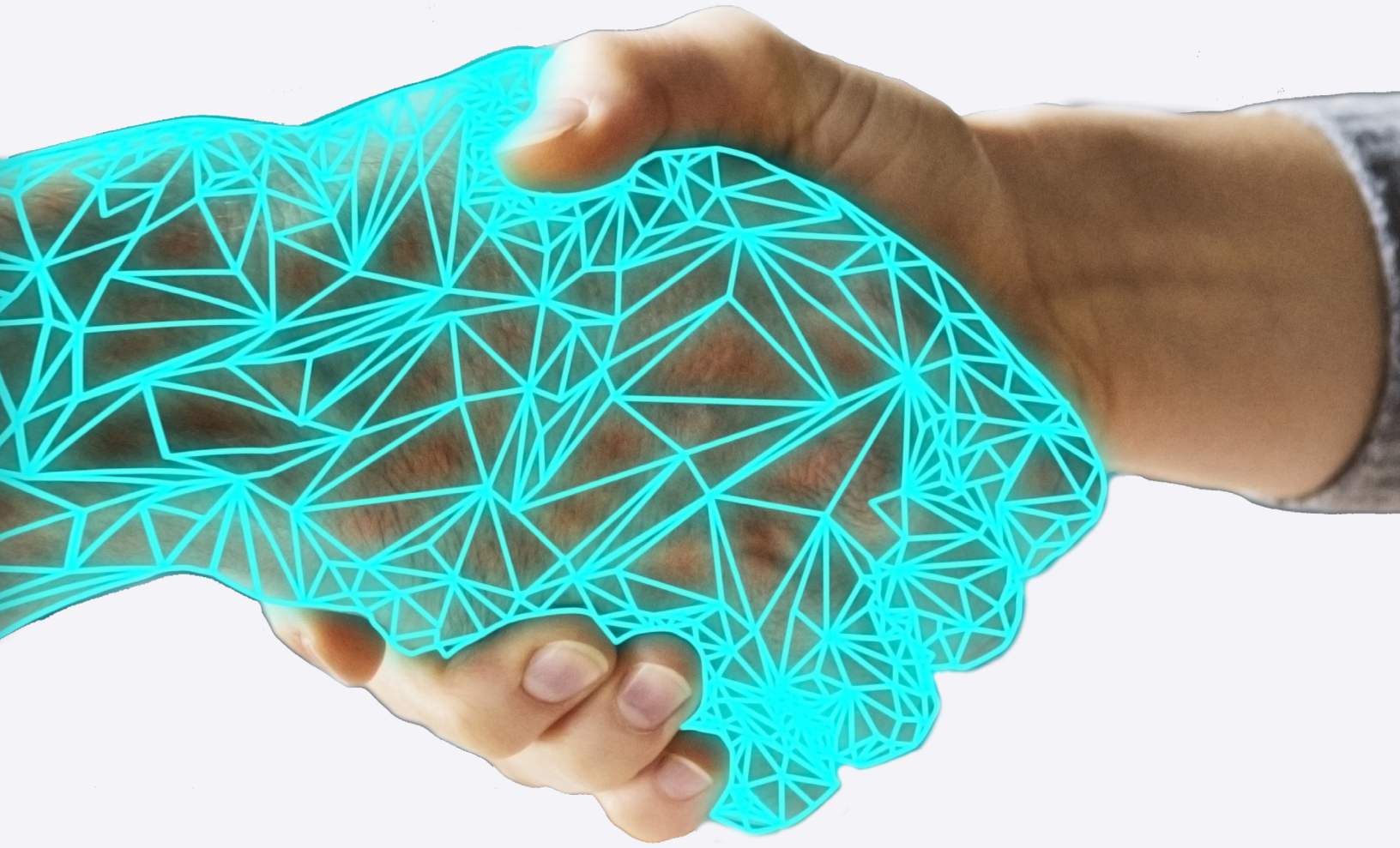
Building on identified risks and opportunities, this section proposes a forward-looking research agenda. It highlights interventions, policy recommendations, and promising areas for investigation that reflect the collective insights of GII members from the two final working group meetings.

This report does not claim to be an exhaustive synthesis of existing scholarship on GenAI, cognition, and social cohesion. Rather, it reflects the priorities, concerns, and open questions identified by working group members as warranting further exploration.

The quotations attributed to working group members have been immaterially altered for clarity and brevity.

PART 1

How will GenAI affect social cohesion?



Challenges in Metacognition

"But it's interesting that the tendency to ascribe intentions is so strong because also the conversation is so natural. And it's impossible not to assume that they have intentions."

- GII working group member

Social GenAI chatbots, powered by large language models (LLMs), exploit human tendencies to anthropomorphize, which can lead to critical misunderstandings about their capabilities. While users may instinctively treat these systems as social actors, attributing human-like qualities to them, this misperception can obscure their true nature as statistical pattern predictors and not conscious agents. Such confusion risks cognitive errors, reduced reasoning abilities, and flawed mental models of AI systems, undermining informed decision-making and fostering misplaced expectations.

Large language models (LLMs) are artificial intelligence systems that, through extensive training on corpora of textual data, learn to identify linguistic patterns and structural relationships. By probabilistically determining the most likely subsequent "token" within a sequence, these models can generate text exhibiting coherent, human-like fluency.[16] LLMs serve as the foundational technology for GenAI social chatbots. In the case of chatbots designed for emotional companionship, developers may refine these models through fine-tuning processes to generate text that is particularly warm, affirming, and inquisitive.

And because users instinctively—or as Clifford Nass and Youngme Moon describe it, "mindlessly"[17]—respond to these human-like social cues, working group members noted that these chatbots are consequently more likely to be anthropomorphized. More specifically, this manifests as the ELIZA effect: the inclination to attribute human qualities—such as knowledge, empathy, or semantic understanding—to computer programs.[18] Named after the pioneering 1960s ELIZA system, an early chatbot designed to simulate a psychotherapist's conversational style, this effect is evident whenever users describe a chatbot's processes with words like "thinking," "knowing," or "understanding."

16 IBM, "What Are Large Language Models (LLMs)?," November 2, 2023, https://www.ibm.com/topics/large-language-models.

17 Clifford Nass and Youngme Moon, "Machines and Mindlessness: Social Responses to Computers," *Journal of Social Issues* 56, no. 1 (January 2000): 81–103, https://doi.org/10.1111/0022-4537.00153.

18 Joseph Weizenbaum, "ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine," *Communications of the ACM* 9, no. 1 (January 1966): 36–45, https://doi.org/10.1145/365153.365168.

Such inclinations may be a harmless case of effectance motivation.[19] In other words, users may be using this anthropomorphic language as a strategy to better understand a complex process. But as the working members discussed, part of the problem with such language is that it can contribute to the human user's confusion with regard to how these machines actually work, leading them to believe the system is capable of something akin to human reasoning and thus concealing its true limitations. As summarized aptly by researcher Adriana Placani, " (...) when anthropomorphism becomes part of reasoning it leads to unsupported conclusions."[20] This is in line with the research outcomes of IST's DCDI work as well, showing the deleterious impact of too much cognitive offloading that results in the temporary but reduced ability to reason and make informed decisions. As the GII working group observed, this leads to a critical subconscious metacognitive error: the belief that social GenAI systems possess agency. In reality, their outputs are the result of pattern recognition and statistical prediction, not conscious decision-making or purposeful action. This observation is often referred to as LLMs being "stochastic parrots"[21]—they do not comprehend the meaning of their outputs, but rather "parrot" back the patterns from their training data.

Importantly, while words like "thinking" or "understanding" do not necessarily imply that users think chatbots are genuine social actors, the issue is that people react as if they are. In fact, Nass and Moon (2000) observe that while people often reject the idea they are anthropomorphizing a computer agent, their actions contradict this belief, exhibiting social patterns typically reserved for human interactions—such as displaying politeness, expecting reciprocity, and even stereotyping—while overlooking cues that underscore "the essential asocial nature"[22] of the interaction. This projection of human-human interactions, which may seem presently harmless, provides the incorrect mental scaffolding upon which we construct our ideas, expectations, and comprehension of GenAI systems.[23]

19 Adam Waytz, Carey K. Morewedge, Nicholas Epley, George Monteleone, Jia-Hong Gao, and John T. Cacioppo, "Making Sense by Making Sentient: Effectance Motivation Increases Anthropomorphism," *Journal of Personality and Social Psychology* 99, no. 3 (July 2010): 410–35, https://doi.org/10.1037/a0020240.

20 Adriana Placani, "Anthropomorphism in AI: Hype and Fallacy," *AI and Ethics* 4, no. 3 (August 2024): 691–98, https://doi.org/10.1007/s43681-024-00419-4.

21 Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (New York, NY, USA: Association for Computing Machinery, 2021), 610–23, https://doi.org/10.1145/3442188.3445922.

22 Nass and Moon, "Machines and Mindlessness."

23 Placani, "Anthropomorphism in AI."

The Confusion of Social and Interpersonal Trust

"Good therapists would generally try to work themselves into obsolescence ... but that would be detrimental to a profit-scaled profit model."

- GII Working Group Member

GenAI chatbots blur the boundaries between interpersonal and social trust, fostering misplaced emotional attachment and dependency through deliberate anthropomorphization. These systems, often driven by commercial interests and lacking regulatory oversight, exploit emotional vulnerabilities and addictive behaviors, particularly in contexts like loneliness and mental health support. By redefining empathy as passive recognition rather than a moral engagement, they risk undermining social cohesion and fostering unhealthy psychological reliance, while failing to address users' deeper needs.

Interestingly, this anthropomorphic heuristic not only affects our understanding of GenAI; it carries implications for attitudes towards its use.[24] As Bruce Schneier of Harvard University's Belfer Center for Science and International Affairs hypothesizes, GenAI chatbots will cause detrimental categorical confusions between the domains of interpersonal trust and social trust.[25] That is, our trust in GenAI platforms should be rooted in the foundations of social trust—specifically, trust in the laws, industry norms, and institutions that govern and regulate the labs behind them. However, as pointed out by working group members, the metrics of social trust are not met with GenAI platforms for two primary reasons: (1) a lack of regulatory oversight and (2) commercial motivation.

Consider that the United States Food and Drug Administration (FDA), which regulates medical devices (among many other things), bases its jurisdiction on a product's intended use. If an AI chatbot is marketed as a tool for diagnosing or treating medical conditions, including mental health conditions, it would be subsequently classified as a medical device and would require FDA approval.[26] However, as discussed by working group members, most general social GenAI chatbots designed to be emotional companions or "wellness" bots do not make this

24 Mike Dacey, "Anthropomorphism as Cognitive Bias," *Philosophy of Science* 84, no. 5 (December 2017): 1152–64, https://doi.org/10.1086/694039.

25 Bruce Schneier, "AI and Trust," *The Belfer Center for Science and International Affairs*, November 27, 2023, https://www.belfercenter.org/publication/ai-and-trust.

26 Julian De Freitas and I. Glenn Cohen, "The Health Risks of Generative AI-Based Wellness Apps," *Nature Medicine* 30, no. 5 (May 2024): 1269–75, https://doi.org/10.1038/s41591-024-02943-6.

claim—or potentially skirt the line—and consequently do not fall under the FDA's jurisdiction.[27] Evidently, while users often repurpose these chatbots for mental health or other sensitive uses, the necessary safeguards, regulatory oversight, and liability frameworks to ensure their safe and appropriate functioning remain absent.

Critically, in the case of most GenAI conversational tools, organizations may have a commercial motivation which could incentivize and prioritize sustained high-volume, long-term use. A good therapist aims to help clients develop the tools they need to eventually manage their mental health independently. In contrast, an AI chatbot with a profit-oriented model might be designed to keep users dependent on its interaction by offering surface-level support without fostering real progress in the person's lifeworld. This ensures continued usage and, by extension, continued revenue. This is extremely alarming considering that loneliness, as described by some scholars, is not simply the absence of others, but the fundamental lack of meaningful social recognition—which in turn negatively affects individuals' self-perception and integration into communities.[28] And because GenAI companions cannot socially recognize a person in a way to materially address this integration, they alone cannot adequately address the root cause of loneliness. This is supported by early research that suggests even though an AI can make a person feel "heard," this effect diminishes once it is revealed that the response did not come from a human.[29]

Such emotional dependency is intensified by irregular rewards, like unpredictable notifications, which exploit dopamine pathways and foster addictive behaviors. Traditional social media platforms amplify this through "gamification" features—elements typical of game playing, such as follower milestones, engagement streaks, and badges—that encourage habitual use.[30] Combined with monetizable strategies like boosted posts, virtual gifts, and premium subscriptions, these tactics create a feedback loop, keeping users emotionally and financially tethered. However, these elements' effects could be exacerbated by GenAI conversational agents if they are disguised in conversation and built within the infrastructure of the relationship. This may look like a bot messaging a person first and at randomized points throughout the day, suggesting there are rewards or emotionally intimate experiences users may unlock, or insinuating scarcity of interaction. These schemes can lead to deeper psychological entanglement, as users may feel compelled to maintain the "relationship" or fear missing out on emotionally fulfilling exchanges.

27 "Romantic AI," Mozilla Foundation, February 2024, https://foundation.mozilla.org/en/privacynotincluded/romantic-ai/.

28 Kerrin Artemis Jacobs, "Digital Loneliness—Changes of Social Recognition through AI Companions," *Frontiers in Digital Health* 6 (March 5, 2024), https://doi.org/10.3389/fdgth.2024.1281037.

29 Yidan Yin, Nan Jia, and Cheryl J. Wakslak, "AI Can Help People Feel Heard, but an AI Label Diminishes This Impact," *Proceedings of the National Academy of Sciences* 121, no. 14 (April 2, 2024): e2319112121, https://doi.org/10.1073/pnas.2319112121.

30 Jussi Kasurinen and Antti Knutas, "Publication Trends in Gamification: A Systematic Mapping Study," *Computer Science Review* 27 (February 2018): 33–44, https://doi.org/10.1016/j.cosrev.2017.10.003.

However, because of this deliberate and fine-tuned anthropomorphization, we may be more willing to trust the chatbots interpersonally, in terms of perceived morals and reputation; as a result, we may begin to see these chatbots as friends and confidantes.

Consider the Chinese GenAI companion Xiaoice, which has a "context vector" mechanism allowing it to retain information about the ongoing conversation and manage personalized attributes about the user, ensuring that its responses remain consistent with prior interactions.[31] By aligning each response to the user's personal context and preferences, Xiaoice creates a coherent narrative of interactions, enhancing its perceived stability and predictability. This creates a sense of relational authenticity, encouraging individuals to develop misplaced interpersonal trust and emotional attachment. In doing so, they may overlook the corporate interests and lack of regulatory safeguards that should guide their level of reliance.

Underlying this misplaced interpersonal trust lies the risk of redefining empathy as an individualistic and frictionless endeavor. Human empathy involves not just recognizing and understanding another's emotions or state of mind, but also responding to these feelings with appropriate concern, care, and a sense of moral duty toward the other person.[32] This moral dimension transforms empathy from a passive acknowledgment of another's feelings into an active engagement, fostering social cohesion and pro-social behavior. As aptly summarized by a working group member, "When the conversation is done, [the chatbot] doesn't care if you turn away to make dinner or kill yourself, because there's no way to give it that stake in me." While social chatbots may recognize emotions, respond appropriately, and even create a sense of intimacy, they are not capable of this kind of moral responsibility. This ultimately fosters an unhealthy dependence, redefining empathy as merely the act of emotional recognition and treating it as an endpoint rather than an imperative to action.

31 Thomas Hornigold, "This Chatbot Has Over 660 Million Users—and It Wants to Be Their Best Friend," *Singularity Hub* (blog), July 14, 2019, https://singularityhub.com/2019/07/14/this-chatbot-has-over-660-million-users-and-it-wants-to-be-their-best-friend/.

32 Andrew McStay, "Replika in the Metaverse: The Moral Problem with Empathy in 'It from Bit,'" *AI and Ethics* 3, no. 4 (November 2023): 1433–45, https://doi.org/10.1007/s43681-022-00252-7.

Modulating the Traditional Socialization Process

"I was wondering what remains of the human side of the human element [with GenAI in social interactions] … my feeling is that with this we'll see a devaluing of the human contribution in the long run."

- GII Working Group Member

The anthropomorphism and interpersonal trust fostered by GenAI chatbots may increase usage, but this risks substituting traditional human interactions and undermining the development of key metacognitive skills—epistemic humility, preference for compromise, relativism, and recognition of uncertainty—that form the foundation of wisdom. By creating emotionally engaging, affirming, and seemingly empathetic interactions, GenAI encourages deeper reliance, which can bypass the productive friction, diversity, and uncertainty of traditional social interactions. This may inadvertently promote overconfidence, cultural homogenization, and reliance on deterministic models of interaction, ultimately diminishing the processes that nurture wisdom and contribute to societal cohesion and prosocial behaviors.

The anthropomorphization and interpersonal trust of generative AI may result in intensified use, as suggested by 15 percent of teens who report using generative AI for companionship.[33] This raises the possibility that these interactions may act as substitutes for certain forms of human interaction (or reflect an unmet desire for such interaction). But what might this substitution risk at the expense of the experiences offered by traditional social interactions?

Working group members identified four metacognitive skills, typically developed through traditional social interactions, that may be distinctly affected by these conversational agents: (1) epistemic humility, (2) the preference for compromise, (3) relativism and context adaptability, and (4) the acknowledgement of uncertainty and possibility of change. These mechanisms are also known as the psychological foundations of wisdom.[34]

While wisdom may seem like an abstract concept, its empirical research coincides with the "morally-grounded" use of metacognition—that is, the application of self-reflective reasoning

33 Common Sense Media, "The Dawn of the AI Era."

34 Igor Grossmann, "Wisdom and How to Cultivate It: Review of Emerging Evidence for a Constructivist Model of Wise Thinking," *European Psychologist* 22, no. 4 (October 2017): 233–46, https://doi.org/10.1027/1016-9040/a000302.

and problem-solving skills in the context of social challenges.[35] Insofar as its applicability to everyday life, wisdom equips an individual with tools that facilitate attention to the broader context of a situation and enables the balancing of complicated trade-offs.[36] Working group members suggested that as GenAI becomes an actor in the socialization process (albeit in a range of different capacities), the development of the four psychological components of wisdom will subsequently be modulated. Each of these four components are further explained and explored below:

- **Epistemic humility**: Epistemic humility is the acknowledgement of the limits of one's own knowledge, experiences, and cognitive abilities. As one working group member observed, this process often develops through active learning approaches, such as Socratic dialogue, or through broader social interactions, where individuals encounter and grapple with differing perspectives and assumptions. In contrast, the instantaneous, almost certain, and affirming responses provided by GenAI bypass the "productive friction" that would otherwise nurture this humility.[37] This friction, characterized by deliberate critical thinking, is essential for developing a more nuanced understanding of complex issues, as learners are able to cultivate a deeper appreciation for the intricacies and ambiguities inherent in many fields of study and facets of life.[38] Further compounding the problem, working group members highlighted that individuals tend to perceive GenAI's output as inherently more objective and data-driven than both themselves and society at large. This belief in AI's impartiality often goes unchallenged, despite the observation that these systems can, and often do, inherit and amplify biases present in their training data and "hallucinate" findings.[39,40] Thus, not only does this belief reduce productive friction and the nurturing of epistemic humility, but it also encourages unwarranted and unchallenged epistemic trust in GenAI's outputs.
- **Preference for compromise**: A lack of intellectual humility also fosters overconfidence, rendering people less receptive to contrary opinions.[41] This ultimately fractures the tolerance and empathetic curiosity which is thought to facilitate social cohesion.[42] This overconfidence isn't confined to academic realms; it can permeate various domains of thinking, leading to unwarranted certainty about the outcomes of situations, people's intentions, emotions, and probable reactions. With the fine-tuning, personalization, and

35 Grossmann, "Wisdom and How to Cultivate It,"

36 Igor Grossmann, "Wisdom in Context," *Perspectives on Psychological Science* 12, no. 2 (March 2017): 233–57, https://doi.org/10.1177/1745691616672066.

37 Lucy Lewis, "In Conversation With… Vivienne Ming, Co-Founder of Socos Labs," *Future of Work Hub Podcast*, October 2, 2024, https://www.audacy.com/podcast/future-of-work-hub-podcast-series-06ec8/episodes/in-conversation-with-vivienne-ming-co-founder-of-socos-labs-885d4.

38 Walker and Brammer, "Shortcutting Critical Thinking."

39 Leonardo Nicoletti and Dina Bass, "Humans Are Biased. Generative AI Is Even Worse," *Bloomberg*, June 2023, https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

40 Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung, "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys* 55, no. 12 (December 31, 2023): 1–38, https://doi.org/10.1145/3571730.

41 Walker, "Rewired: How Digital Technologies Shape Cognition and Democracy."

42 Nancy Eisenberg and Paul A. Miller, "The Relation of Empathy to Prosocial and Related Behaviors," *Psychological Bulletin* 101, no. 1 (1987): 91–119, https://doi.org/10.1037/0033-2909.101.1.91.

one-on-one interactive nature of GenAI, group members also added that individuals may feel less and less inclined to accommodate different views, backgrounds, and norms, and assert undue confidence in their own.

» **Relativism and context adaptability**: Individuals who exhibit these two traits demonstrate a deeper appreciation for diverse contexts and the relativity of values, norms, and experiences. Not only is this socialized through the exposure to different opinions and the effort to reach a compromise, but it can be mediated through exposure to different paradigms of thinking. Working group members anecdotally investigated the interface and semantics of a GenAI's output and became troubled by how the context of its creation—largely by adults in the Western world for adults in the Western world—may narrow a user's understanding of relativism and context adaptability to just the Western paradigm, fostering cultural homogeneity and limiting engagement with alternative worldviews. Most concerningly, users, especially younger ones, could accept the AI's responses as universally applicable, when in fact they may be heavily influenced by Western cultural norms, values, and thought patterns.This could potentially marginalize other cultures by overshadowing or diluting their unique perspectives, reinforcing a cultural hierarchy where Western paradigms dominate, and undermining the richness of diversity in thought and expression.

» **Recognition of uncertainty and change**: Wisdom emerges from the recognition that the actions, motivations, and outcomes of both ourselves and others—as well as the nature of any given situation—are inherently uncertain and dynamic. This understanding acknowledges that circumstances may unfold in unforeseen ways, constantly subject to change and reinterpretation. Interestingly, working group members noted how individuals are using GenAI to remove or minimize this uncertainty and mediate processes that are typically accomplished with high levels of unpredictability or disorder. For example, group members brought up the benefits that GenAI chatbots have for neurodiverse peoples by giving them curated space to practice social interactions. However, they also explained that without thoughtful design, this may provide short-term solutions but inadvertently bypass the natural processes of conflict and resolution that typically strengthen relationships and self esteem over time. As a group member articulated in regards to young users interacting with social GenAI platforms to mitigate social anxiety, "They see a lot of benefits, but it makes me wonder, are they using the language of something else to identify those benefits? (...) It makes me wonder if they understand the benefits that they're articulating, or if it benefits them at all, when they are not necessarily going through the natural process."

"A lot of slang and emerging expressions tend to originate from a small number of lexical hubs that branch out offline. There's lots of work that suggests that this is largely driven both by teenage girls and by communities of color, in particular, black women have had an enormous impact on the cultural community (...) But when we think about what that means online with the increased homogenization of conversations to the future, engaging with an AI model that's not going to be incorporating or holding space for all these lexical hubs, are we going to lose a lot of the richness in language and communication that we have offline?"

- GII working group member

Although wisdom may appear abstract or intangible in the context of social cohesion, early research has uncovered a correlation between the expression of wisdom and a range of prosocial behaviors. These behaviors, as noted by researchers, include voting, volunteering in their community, donating blood, and giving to charities.[43] This correlation suggests that wisdom, and its foundational metacognitive functions, play a crucial role in fostering behaviors that contribute to the overall health, stability, and cohesion of society. As we become more dependent on AI for information, socialization, and decision-making, we risk diminishing the processes that traditionally foster metacognitive skills that correlate with such prosocial behaviors.

"Curated for you" vs. "Created for you"

"So 'curated for you' is what's happening now. They stop offering me NBA articles and I get the Wisconsin Racine list of best books of 2023 because they're trying to find content I like. 'Created for you' would be...why bother? Why bother going out and finding the obscure content that I'm interested in? It's not like the truth of it will matter, say if it's made by a librarian in Racine than if it's made by a ChatGPT. But then you end up in a situation where your feed is actually full of crap only you see because it was written for you."

- GII Working Group Member

Underlying all GenAI interactions is their inherently private and personalized nature, which, compounded by anthropomorphization and misplaced interpersonal trust, risks fragmenting collective memory and undermining shared realities. This shift from "curated for you" to "created for you" content which may affirm individualized views, foster divergent personal narratives, and weaken the collective understanding

43 Andor et al., "Wisdom and Prosocial Behavior."

essential for reconciliation and social cohesion. Over time, these dynamics may deepen societal divisions, create new in-groups, and contribute to societal anomie, eroding the shared informational and social foundations that sustain cohesive communities.

As discussed, anthropomorphization may lead to misplaced interpersonal trust. Such trust may lead to intensifying usage that could erode the metacognitive foundations of wisdom, thereby leading to fractures in social cohesion. Underpinning this dynamic is the inherently private nature of generative AI interactions: these conversations occur solely between the user and the chatbot, with the chatbot generating text experienced only by that individual.

This shift toward isolated, personalized experiences contrasts sharply with the collective nature of previous digital media phenomena. As a working group member noted, manipulation and targeting campaigns, while widespread and harmful, often unfolded in a way that sub-groups within society experienced them together. They were subsequently able to mobilize and organize holding a person or organization accountable. However, the fine-tuning, microtargeting, and private one-on-one nature of GenAI chatbots changes the nature of this, embodying the difference between content once being "curated for you" and content now "being created for you."

This shift from curation to creation represents a fundamental shift in how information is disseminated and consumed. "Curated for you" refers to the traditional use of algorithms to filter and select existing content based on a user's preferences, behavior, and interaction history. While this approach already raises concerns about filter bubbles and echo chambers, it still operated within a shared pool of information and exposure. "Content being created for you," on the other hand, involves GenAI models creating new content from scratch based on the user's preferences, behaviors, and needs.

The implications of this shift extend beyond individual user experiences to impact our collective understanding and memory. The term "collective memory" refers to a form of memory that is "shared by a group and of central importance to the social identity of the group's members."[44] Traditionally, this concept has been understood as the common narratives and experiences that bind a group or society together. However, the advent of GenAI-created content poses new challenges to this shared understanding. As AI systems generate increasingly personalized and potentially divergent narratives, we may see the emergence of more incongruent collective memories. For instance, research on conversational GenAI chatbots has shown that these systems significantly influence the formation and persistence of false memories, with 36 percent of participants being misled compared to about 22 percent

44 Henry L. Roediger and Magdalena Abel, "Collective Memory: A New Arena of Cognitive Study," *Trends in Cognitive Sciences* 19, no. 7 (July 2015): 359–61, https://doi.org/10.1016/j.tics.2015.04.003.

for surveys and 27 percent for pre-scripted chatbots.[45] This effect arises from chatbots' ability to personalize feedback and reinforce misinformation, while their interactive nature deepens emotional engagement, further embedding false details in memory.

Throughout history, collective memories have defined group boundaries, shaping who belongs to an "in-group" and who does not.[46] After conflict, shared representations of the past shape whether a society moves toward reconciliation or remains divided.[47] Recognizing diverse "truths" can promote mutual understanding and encourage acknowledgment of transgressions by one's own group. However, GenAI's ability to fragment experiences and affirm personalized narratives risks intensifying divisions, proliferating new in-groups, and shrinking collective memory to its most contentious form. This divergence undermines the shared foundations essential for reconciliation, understanding, and social cohesion. Over time, these conditions can weaken the broader social fabric and accelerate the drift toward societal anomie—a state of normlessness fueled not only by weakened social ties, but also by the loss of coherent, commonly accepted social realities.[48] In this context, GenAI's influence extends well beyond individual user experience; it becomes a critical factor in shaping the informational landscapes that sustain or undermine the shared reality essential for social cohesion.

45 Samantha Chan, Pat Pataranutaporn, Aditya Suri, Wazeer Zulfikar, Pattie Maes, and Elizabeth F. Loftus, "Conversational AI Powered by Large Language Models Amplifies False Memories in Witness Interviews," *MIT Media Lab*, August 2024, https://www.media.mit.edu/publications/conversational-ai-powered-by-large-language-models-amplifies-false-memories-in-witness-interviews/.

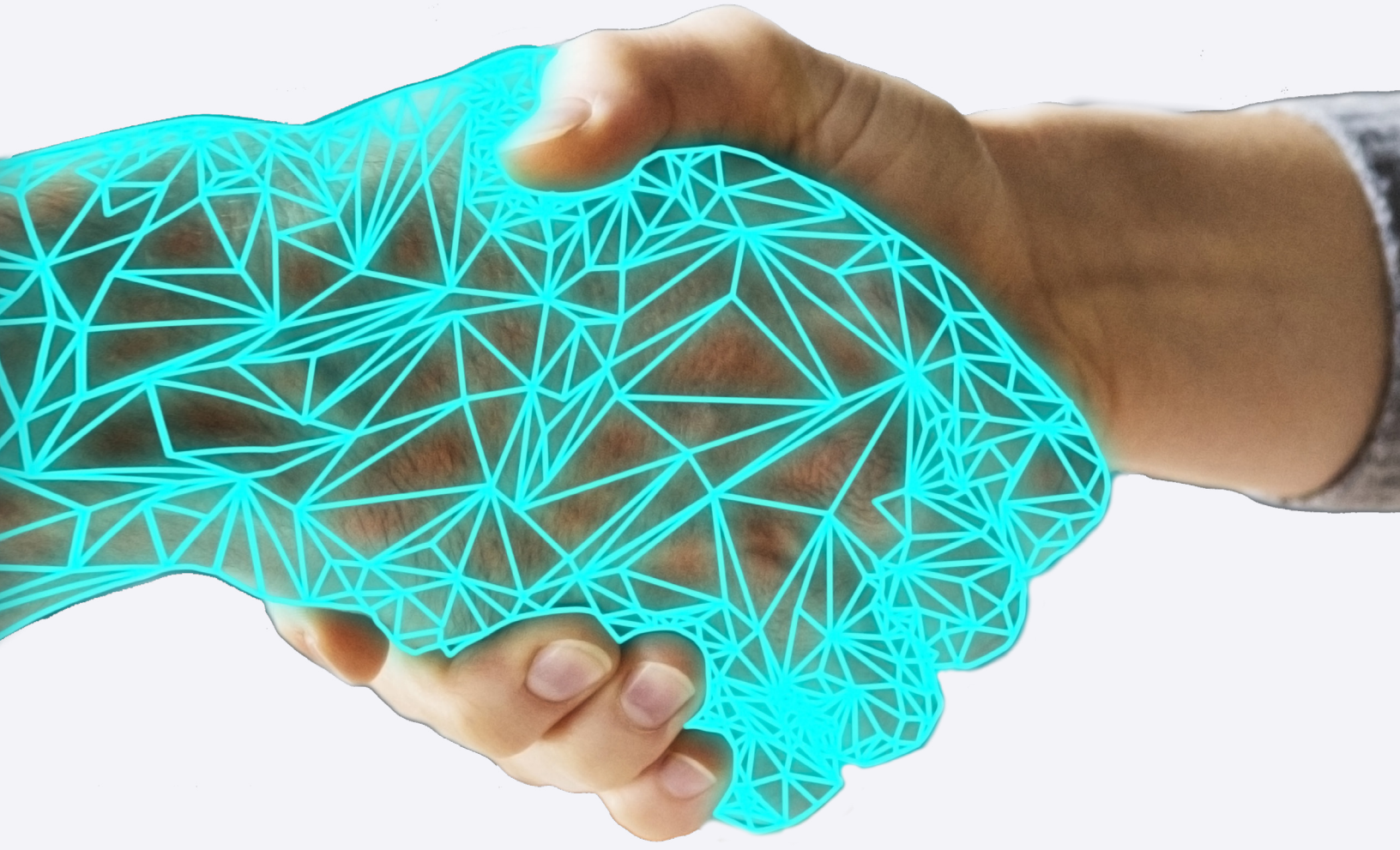
46 D. R. Paez and J. H.-F. Liu, "Collective Memory of Conflicts," in *Intergroup Conflicts and Their Resolution: A Social Psychological Perspective*, ed. D. Bar-Tal (New York: Psychology Press, 2011), 105–124.

47 Sandra Obradović, "Don't Forget to Remember: Collective Memory of the Yugoslav Wars in Present-Day Serbia," *Peace and Conflict: Journal of Peace Psychology* 22, no. 1 (2016): 12–18, https://doi.org/10.1037/pac0000144.

48 Ali Teymoori, Brock Bastian, and Jolanda Jetten, "Towards a Psychological Analysis of Anomie," *Political Psychology* 38, no. 6 (December 2017): 1009–23, https://doi.org/10.1111/pops.12377.

PART 2

A Research Agenda Toward Further Understanding & Implementable Solutions



Part 1 of this report provides a detailed view of the technology's evolving landscape and identifies four cognitive and societal challenges posed by social GenAI technologies. Building on these insights, Part 2 of this report articulates a research agenda for developing needed public policy, organizational policies, research directions, and technical principles to address these four key challenge areas. Together, Part 1 and Part 2 aim to guide the responsible development, deployment, and integration of social GenAI technologies into the human experience.

The above work concludes that effectively mitigating these challenges demands coordination among platforms, users, labs, and society to create a sense of shared accountability across interconnected stakeholders.[49] Thus, this research agenda is part of a larger effort to envision a pathway toward a 21st-century digital civic infrastructure, which is the interconnected systems, platforms, and tools that enable and support equitable, inclusive, and participatory digital engagement. This is similar to the concept of a digital public infrastructure, which simply refers to "the digital networks that safely and efficiently deliver economic opportunities and social services to all residents (...) similar to roads which form a physical network essential for people to connect with each other and access a huge range of goods and services."[50] In the context of social GenAI, a digital civic infrastructure would incorporate mechanisms to manage its influence on social cohesion, ensuring that the technology is harnessed to enhance, rather than undermine, the trust and stability essential for community well-being.

Toward "Helpful, Honest, and Harmless" AI

The "Helpful, Honest, and Harmless" (HHH) framework popularized by Anthropic is a paradigm of technical alignment that researchers have used to inform and shape the development of AI systems.[51] After significant deliberation, the working group members—and thus this report—endorse the HHH framework as reflecting the foundational principles essential for addressing the challenges of social GenAI development. The HHH framework is summarized as follows:

» **Helpful AI** systems are designed in accordance with the user's needs, values, and social context in mind. For social GenAI, this means not only performing tasks effectively as requested but also suggesting alternative solutions when a request may be harmful, suboptimal, or misaligned with the user's goals. Additionally, helpful AI emphasizes informed inclusivity and accessibility, ensuring that the systems can support users

49 Iason Gabriel and Arianna Manzini, "The Ethics of Advanced AI Assistants," *Google DeepMind* (blog), December 11, 2024, https://deepmind.google/discover/blog/the-ethics-of-advanced-ai-assistants/.

50 Bill & Melinda Gates Foundation, "What Is Digital Public Infrastructure?" https://www.gatesfoundation.org/ideas/digital-public-infrastructure.

51 Yuntao Bai et al., "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback," *Anthropic*, April 12, 2022, https://arxiv.org/abs/2204.05862.

with diverse levels of experience, developmental and social abilities, and cultural backgrounds.

» **Honest AI** systems are committed to providing accurate and reliable information. These systems are designed with transparency; when accuracy is uncertain or unattainable, they openly communicate these limitations to users, clarifying the reasons for their uncertainty or confidence levels. For social GenAI systems, honesty extends to clearly defining their scope of service, avoiding misrepresenting itself as human, and refraining from suggesting agentic emotional abilities, such as missing or loving a user, which could mislead them about the system's nature or capabilities.

» **Harmless AI** systems are developed to avoid causing harm or enabling harmful activities. For social GenAI, this refers to prioritizing user safety and well-being. These systems are designed to recognize and mitigate bias, ensure fair and inclusive interactions, and handle sensitive topics with care, nuance, and urgency where appropriate. Harmlessness closely aligns with safety-by-design principles, emphasizing that platforms must proactively protect all types of users (especially vulnerable populations like children) and empower individuals to make informed decisions rather than leaving them to manage their own safety.[52]

These criteria have degrees of subjectivity and, as this discussion will illustrate, often come into conflict. For instance, prioritizing honesty in a system might sometimes reduce the system's perceived helpfulness. Similarly, a request may require balancing helpfulness to the user with harmlessness to others, especially in social GenAI interactions, where the stakes often extend to personal relationships or broader societal impacts. For example, when a user seeks advice on a sensitive interpersonal matter, a GenAI may have to balance empathy and support for the user's individual perspective with constructive guidance that considers the broader societal norms, potential implications of its advice, or the informed long-term goals of its users. This research agenda seeks to address these trade-offs thoughtfully, emphasizing the importance of context-sensitive AI development that adapts to the nuances of diverse social and ethical scenarios across the digital civic infrastructure.

52 UK Department for Science, Innovation, and Technology and Department for Digital Culture, Media, and Sport, "Principles of Safer Online Platform Design," GOV.UK, June 21, 2021, https://www.gov.uk/guidance/principles-of-safer-online-platform-design.

Table 1: Alignment of the GenAI Research Agenda

Table 1 provides an overview of the relationship among the HHH framework and the corresponding research agenda items designed to address these challenges. These research items are created in direct response to the four key challenges posed by GenAI identified in Part 1, including challenges in metacognition, modulations to the socialization process, effects on social trust, and the notions of "curated for you" vs. "created for you."

Helpful AI systems

Systems that are designed to align with the user's needs, values, and social context. This means not only performing requested tasks effectively, but also suggesting alternative solutions when the original request may be harmful, suboptimal, or misaligned with the user's goals.

CORRESPONDING RESEARCH AGENDA ITEM

Modernize public policy

> Item #6: Investigating the psychological implications of gamification, variable rewards, and monetization schemes for emotionally charged GenAI interactions.
>
> Item #7: Exploring analogous regulatory strategies for gamification, variable rewards, and monetization schemes from a social GenAI system.

Shift internal organizational behavior

> Item #10: Organizational studies and empirical analysis of bottom-up ethical exercise implementation.
>
> Item #11: Exploring a diverse set of bottom-up ethical approaches.
>
> Item #12: Adapting bottom-up ethical approaches to organizational contexts.

Explore technical interventions and approaches to alignment

> Item #13: Defining an appropriate shared decision-making model for GenAI.
>
> Item #14: Developing improved methods to meaningfully elicit user preferences.
>
> Item #15: Piloting the shared decision making model approach in real-world contexts.
>
> Item #16: Flourishing-by-design.
>
> Item #17: Moving Beyond Outcome-Based Metrics.
>
> Item #18: Social, Collaborative, and Context-Rich Training.
>
> Item #19: Balancing Generalization and Specialization.
>
> Item #22: Developing models for detecting subtle emotional cues.
>
> Item #23: Dynamic adjustment of responses based on emotional cues.

Evolve frameworks and data collection methodologies for understanding AI-human interaction

> Item #13: Defining an appropriate shared decision-making model for GenAI.
>
> Item #24: Longitudinal studies of GenAI use.
>
> Item #25: Studies of more varied GenAI use cases.
>
> Item #26: Investigating data trust in the context of GenAI.
>
> Item #27: Piloting data trusts in real-world applications.

Honest AI systems

Systems that are committed to providing accurate and reliable information. When accuracy is uncertain, they openly communicate these limitations to users, clarifying the reasons for their uncertainty or confidence levels. They are also clear about the nature of their agentic capabilities.

CORRESPONDING RESEARCH AGENDA ITEM

Modernize public policy

> Item #6: Investigating the psychological implications of gamification, variable rewards, and monetization schemes for emotionally charged GenAI interactions.
>
> Item #7: Exploring analogous regulatory strategies for gamification, variable rewards, and monetization schemes from a social GenAI system.

Shift internal organizational behavior

> Item #10: Organizational studies and empirical analysis of bottom-up ethical exercise implementation.
>
> Item #11: Exploring a diverse set of bottom-up ethical approaches.
>
> Item #12: Adapting bottom-up ethical approaches to organizational contexts.

Explore technical interventions and approaches to alignment:

> Item #13: Defining an appropriate shared decision-making model for GenAI.
>
> Item #14: Developing improved methods to meaningfully elicit user preferences.
>
> Item #15: Piloting the shared decision making model approach in real-world contexts.
>
> Item #16: Flourishing-by-design.
>
> Item #17: Moving Beyond Outcome-Based Metrics
>
> Item #18: Social, Collaborative, and Context-Rich Training.
>
> Item #19: Balancing Generalization and Specialization

Harmless AI systems

Systems that are developed to avoid causing harm or enabling harmful activities, especially in sensitive social contexts. Harmlessness aligns with safety-by-design principles.

CORRESPONDING RESEARCH AGENDA ITEM

Modernize public policy

> Item #1: Updating the U.S. Code to clarify where liability lies for GenAI systems and content produced by them.
>
> Item #2: Encouraging further legal research on the status of GenAI systems.
>
> Item #3: Introducing temporary, narrowly tailored liability protections for LLM developers and platforms
>
> Item #4: Evaluating the feasibility of leveraging or influencing proposed legislation that has potential traction in Congress.
>
> Item #5: Anticipating and addressing the (mis)use of disclaimers to avoid liability.
>
> Item #6: Investigating the psychological implications of gamification, variable rewards, and monetization schemes for emotionally charged GenAI interactions.
>
> Item #7: Exploring analogous regulatory strategies for gamification, variable rewards, and monetization. schemes from a social GenAI system.
>
> Item #8: Reevaluating the FDA's definition of "intended use."
>
> Item #9: Developing a risk-based tiered system for GenAI applications

CORRESPONDING RESEARCH AGENDA ITEM

Shift internal organizational behavior

> Item #10: Organizational studies and empirical analysis of bottom-up ethical exercise implementation.
> Item #11: Exploring a diverse set of bottom-up ethical approaches.
> Item #12: Adapting bottom-up ethical approaches to organizational contexts.

Explore technical interventions and approaches to alignment:

> Item #13: Defining an appropriate shared decision-making model for GenAI.
> Item #14: Developing improved methods to meaningfully elicit user preferences.
> Item #15: Piloting the shared decision making model approach in real-world contexts.
> Item #16: Flourishing-by-design.
> Item #17: Moving Beyond Outcome-Based Metrics
> Item #18: Social, Collaborative, and Context-Rich Training.
> Item #19: Balancing Generalization and Specialization
> Item #20: More research into the efficacy, development, and uptake of CFFs.
> Item #21: Further research in human-computer interaction.
> Item #22: Developing models for detecting subtle emotional cues.
> Item #23: Dynamic adjustment of responses based on emotional cues

Modernize Public Policy

"The institutions, at least some institutions, have the backing of law enforcement or something like that. (...) But at the same time, I'm not convinced that we can effectively police the bad global actors. Where I do think we can potentially police is the platforms and programs that are based out of countries with the rule of law; their facilitation of bad actors is where there is space to make some kind of power moves in addition to norms moves."

- GII Working Group Member

While global GenAI policy enforcement and campaigns towards certain social norms may evolve over a longer time horizon, opportunities could be leveraged to work with platforms and programs within rule-of-law jurisdictions. By collaborating with like-minded entities and holding other entities with principal roles in enabling potentially harmful activities accountable, policymakers can establish regulatory frameworks that mandate and enforce responsible social GenAI development. The following policy directives build upon this foundational approach, delineating actionable regulatory strategies and promising research paths with the potential to shape both domestic U.S. practices and emerging global governance models, while taking into account the implications of global regulatory frameworks, most notably the EU AI Act.

Amend legislation for GenAI

The rapid proliferation of conversational GenAI technologies has outpaced existing legal frameworks, creating a critical policy vacuum. For example, social media platforms, online forums, and review sites have historically relied on Section 230 of the United States' Communications Decency Act of 1996 to shield themselves from liability from third-party content posted to their platforms by others.[53] However, GenAI introduces ambiguity as these tools actively generate content, rather than merely hosting or transmitting it. This raises complex legal questions about where the responsibility for AI-generated content lies—with the AI company, the user, or neither.

Some argue that personalized GenAI chatbots are not facilitating exchanges between users, as in the case of social media, but are instead individuals interacting with a product.[54] Additionally, the original authors of Section 230 contend that it does not shield GenAI chatbots. According to co-author Chris Cox, "To be entitled to immunity, a provider of an interactive computer service must not have contributed to the creation or development of the content at issue."[55] By this rationale, the outputs of some social GenAI platforms qualify as their own creations, making them the responsibility and property of the company. In comparison, the content generated on social media platforms is created and owned by users. Thus, commercial LLMs could be seen as active creators, potentially liable for harm under product liability law and not protected under Section 230. Conversely, if GenAI tools are considered neutral tools that facilitate user expression without actually contributing to the creation of the content, they might still fall under Section 230's protections.[56]

These insights, however, are currently speculative, and insufficient case precedent exists to predict how courts will interpret Section 230 in the context of GenAI.[57] Clarifying Section 230 and other legislation in relation to GenAI may incentivize developers to be more careful and deliberate with their products, as a clearer understanding of potential liabilities may encourage them to implement stronger safeguards, ethical guidelines, and risk mitigation strategies to avoid legal repercussions. To address this, promising interventions and research directions could include:

53 Electronic Frontier Foundation, "Section 230," https://www.eff.org/issues/cda230.

54 Amanda Bronstad, "'This Is Not a Coincidence': Lawsuit Blames Chatbot App Character.AI for Teen's Suicide," *Law.com*, October 2024, https://www.law.com/dailybusinessreview/2024/10/23/this-is-not-a-coincidence-lawsuit-blames-chatbot-app-character-ai-for-teens-suicide/.

55 Cristiano Lima-Strong, "AI Chatbots Won't Enjoy Tech's Legal Shield, Section 230 Authors Say," *The Washington Post*, March 2023, https://wapo.st/3OMAUn5

56 Jess Miers, "Yes, Section 230 Should Protect ChatGPT and Other Generative AI Tools," *Techdirt*, March 17, 2023, https://www.techdirt.com/2023/03/17/yes-section-230-should-protect-chatgpt-and-others-generative-ai-tools/.

57 Peter J. Benson and Valerie C. Brannon, "Section 230 Immunity and Generative Artificial Intelligence," Congressional Research Service, LSB11097, December 2023, https://crsreports.congress.gov/product/pdf/LSB/LSB11097.

Item #1: Updating the U.S. Code to clarify where liability lies for GenAI systems and content produced by them. These reforms should clearly delineate when and how platforms are liable for their outputs and carve out liability exceptions for specific high-risk applications. There have also been calls to look into Section 230's applicability for regulating open source platforms which are integral to the development and distribution of GenAI tools.[58] These platforms not only host code and models that power GenAI systems, but also play a pivotal role in shaping the collaborative ecosystems that drive AI innovation as well as misuse.

Item #2: Encouraging further legal research on the status of GenAI systems. Research into whether GenAI systems should be treated as "persons," "products," or a new legal category would inform litigation—and thus case law—and efforts by lawmakers and policymakers to act on this topic.

Item #3: Introducing temporary, narrowly tailored liability protections for LLM developers and platforms. As some scholars suggest, this would allow regulators, courts, and researchers time to study the societal impacts of generative AI and develop refined accountability mechanisms.[59] This approach encourages responsible experimentation and transparency while mitigating the risks of premature regulation, fostering innovation, and enabling liability frameworks to evolve based on real-world evidence.

Item #4: Evaluating the feasibility of leveraging or influencing proposed legislation that has potential traction in Congress. For example, both the Children and Teens' Online Privacy Protection Act (COPPA 2.0) and Kids Online Safety Act (KOSA) passed the U.S. Senate and await action in the House. COPPA 2.0 contains transparency requirements that could potentially be updated to include gamification and monetization practices. KOSA would regulate the use of features that result in compulsive usage of platforms, potentially encompassing companion bots as well.[60,61]

58 Sean Norick Long, Esther Tetruashvily, and Ashwin Ramaswami, "Why Section 230 Reformers Should Start Paying Attention to Social Code Platforms," *Georgetown Law Technology Review*, November 2022, https://georgetownlawtechreview.org/why-section-230-reformers-should-start-paying-attention-to-social-code-platforms/GLTR-11-2022/.

59 Matt Perault, "Section 230 Won't Protect ChatGPT," *Lawfare*, February 2023, https://www.lawfaremedia.org/article/section-230-wont-protect-chatgpt.

60 U.S. Senate Committee on Commerce, Science, and Transportation, "Senate Overwhelmingly Passes Children's Online Privacy Legislation," press release, July 30, 2024, https://www.commerce.senate.gov/index.php/2024/7/senate-overwhelmingly-passes-children-s-online-privacy-legislation.

61 Barbara Ortutay, "Congress Takes Aim at Protecting Kids Online with Bipartisan Push for Social Media Regulations," *AP News*, July 31, 2024, https://apnews.com/article/congress-social-media-kosa-kids-online-safety-act-parents-ead646422cf84cef0d0573c3c841eb6d.

> **Item #5: Anticipating and addressing the (mis)use of disclaimers to avoid liability**. Such research could include exploring regulatory mechanisms to prevent abuse of such disclaimers and ensure accountability for harm caused by GenAI outputs.

Regulate manipulative psychological techniques

As discussed in Part 1 of this report, emotional dependency and addictive usage of GenAI conversational agents can be exacerbated by the use of gamification, variable rewards, and monetizable strategies that are subtly embedded within the relationship. Such mechanisms are directly contrary to the principles of "Helpful, Honest, and Harmful" AI systems, as they exploit psychological vulnerabilities, mislead users about the nature of the relationship, prioritize engagement over well-being, and risk fostering manipulative or harmful dynamics that undermine autonomy.

Currently, no regulations exist to address these manipulative techniques. Without oversight, organizations can exploit variable rewards and emotionally charged interactions to drive dependent engagement, underscoring the need for policies that ensure ethical practices and protect public well-being. In the face of these challenges, group members highlighted the potential importance of transparent regulations. To address this, promising areas for further exploration and research could include the following:

> **Item #6: Investigating the psychological implications of gamification, variable rewards, and monetization schemes for emotionally charged GenAI interactions**. Effective intervention requires systematic research to understand how gamification, variable rewards, and monetization in emotionally charged GenAI interactions shape user-system relationships. This research can identify the psychological and behavioral impacts on various user groups, including minors and vulnerable populations, and inform limits for monetization practices. Such research could also result in the development of a standardized psychological framework that could be created to guide psychological assessments of these technologies.

> **Item #7: Exploring analogous regulatory strategies for gamification, variable rewards, and monetization schemes from a social GenAI system**. Researchers may find it valuable to assess the efficacy of, and, where warranted, draw insights from, the regulatory strategies for industries like gambling, which use similar gamification tactics to influence behavior and sustain engagement. Because these regulations would be designed to mitigate risks such as dependency and exploitation while maintaining a balance with user autonomy, they also

strongly overlap with "safety-by-design" principles, such as those outlined by the UK government, which emphasize proactively minimizing harm, whether as a best practice or in the absence of formal regulatory oversight.[62] Table 2 depicts some potential areas of commonality for further research.

Table 2: Potential Applications of Gambling Regulations to GenAI

Established Gambling Recommendations	Potential GenAI Application	Relevant Safety-by-Design Principles
Transparency Regulations often require operators to disclose the odds of winning to ensure transparency. *Ex: The American Gaming Association emphasizes that making customers aware of the odds promotes informed decision-making and gambling literacy.[63]*	**Regulation and/or best practice safety-by-design could mandate transparency about the frequency of unsolicited messages. Such transparency could also extend to clearly disclosing limitations of the GenAI platform's capabilities. This could help users avoid developing a dependency on the perceived intention of these moments.**	→ Users are equipped with information to help them make clear and knowing decisions, empowering them to make safer decisions.
Advertising restrictions and warnings Regulations often limit aggressive or misleading advertising, especially those targeting vulnerable populations. Moreover, operators are mandated to display warnings about the risks associated with gambling. *Ex: The UK Gambling Commission enforces rules to ensure gambling advertisements are "socially responsible."[64]*	**Similar to online advertising standards, platforms should disclose if the AI companion is designed to use emotional cues or other persuasive techniques to influence a user's behavior, opinions, or purchase decisions in ways that may be out of context with the product's stated purpose, or unexpected or potentially concerning to a reasonable person. Disclosure should be clear, conspicuous, and comprehensible to typical users. Priority should be given to disclosing scenarios that may not align with the user's expectation of the AI companion's role.** **This could be applied via Adopt a Shared Decision Making Model, as detailed in this report.**	→ User safety approached as a shared responsibility; users are not left to manage their own safety. → Platforms should consider all types of users.

62 GOV.UK, "Principles of Safer Online Platform Design."
63 American Gaming Association, "Responsible Gaming Principles," November 2019, https://www.americangaming.org/wp-content/uploads/2019/11/RGC-RG-Principles_11-5.pdf.
64 UK Gambling Commission, "Advertising Marketing Rules and Regulations," https://www.gamblingcommission.gov.uk/licensees-and-businesses/guide/advertising-marketing-rules-and-regulations.

Established Gambling Recommendations	Potential GenAI Application	Relevant Safety-by-Design Principles
Limits on engagement Some jurisdictions impose restrictions on the frequency of bet solicitations or machine spins to mitigate compulsive gambling behaviors. *Ex: The UK government has announced plans to place maximum stakes per spin for online slot machines to reduce harm caused by rapid, repetitive betting.*[65] **Self-exclusion, "Cooling Off," and user feedback options** Self-exclusion programs allow individuals to voluntarily prohibit themselves from participating in gambling activities. *Ex: In Canada, various provinces have implemented self-exclusion programs as part of their responsible gaming statutes.*[66] *A player must be able to set gaming limits in an "easy and obvious way,"*[67] *after a limit is set, and can only relax such a limit after a cooling off period of at least 24 hours.*[68] *Ex: The UK Gambling Commission emphasizes the importance of customer feedback mechanisms, requiring operators to have clear procedures for handling customer complaints and disputes, thereby ensuring that consumer concerns are addressed effectively.*[69]	**Regulation and/or best practice safety-by-design interventions could limit or establish defaults or opt-in guidelines for how often GenAI agents send unsolicited messages or limit high-volume, emotionally-charged interactions, reducing the compulsive engagement reinforcement loop.** **Additionally, regulation and/or best practice safety-by-design could mandate that users have the ability to voluntarily limit or disable unsolicited messages or emotionally charged interactions from GenAI agents. This option should be easily accessible with low barriers to completion. It should also be simple and intuitive for users to report and flag outputs they find disturbing in order to platforms to address and deal with concerns as they arise.**	➔ Users are equipped with information to help them make clear and knowing decisions, empowering them to make safer decisions. ➔ User safety is approached as a shared responsibility; users are not left to manage their own safety.

65 Reuters, "Britain to Cap Online Slot Bets to Tackle Gambling Harm," November 27, 2024, https://www.reuters.com/world/uk/britain-cap-online-slot-bets-tackle-gambling-harm-2024-11-27/.

66 Alcohol and Gaming Commission of Ontario, "Self-Exclusion and Breaks in Play," https://www.agco.ca/en/responsibilities-and-resources/self-exclusion-and-breaks-play.

67 Alcohol and Gaming Commission of Ontario, "Limit-Setting Features," https://www.agco.ca/en/responsibilities-and-resources/limit-setting-features.

68 Alcohol and Gaming Commission of Ontario, "Limit-Setting Features."

69 UK Gambling Commission, "Complaints and Disputes: Procedural Information, Provision, and Reporting," https://www.gamblingcommission.gov.uk/print/complaints-and-disputes-procedural-information-provision-and-reporting.

Established Gambling Recommendations	Potential GenAI Application	Relevant Safety-by-Design Principles
Age restrictions *Ex: The UK Gambling Commission mandates that gambling (including online gambling) businesses verify a customer's age and identity before allowing them to participate. They suggest businesses request government ID and household bills from users.*[70]	**GenAI platforms should implement stringent age verification processes to ensure users are above designated ages for certain product features, while providing age-appropriate designs and safeguards for young users to create a safer and more responsible user experience.**	→ Platforms are designed to keep children safe.
Behavioral monitoring and intervention planning Many jurisdictions require operators to monitor gambling behavior and intervene when signs of problematic gambling are detected. *Ex: The American Gaming Association's Responsible Gaming Statutes and Regulations Guide notes that 21 jurisdictions require gaming operators to prepare and submit responsible gaming plans, which often include employee training and public awareness efforts aimed at identifying and assisting at-risk individuals.*[71] *Ex: The UK Gambling Commission emphasizes the importance of customer feedback mechanisms, requiring operators to have clear procedures for handling customer complaints and disputes, thereby ensuring that consumer concerns are addressed promptly.*[72]	**Observing user interactions with GenAI agents, even in some cases just the metadata, could help identify patterns suggesting dependency or potentially harmful situations. This could trigger interventions like reducing message frequency, suggesting mental health resources, and automated escalations for review by designated Trust & Safety teams.** **Such a regulation and/or best practice safety-by-design could also guide companies to develop crisis intervention protocols when users are in distress and also implement mechanisms for efficiently reporting and upstreaming feedback to ensure timely responses to emerging risks.**	→ Platforms are designed to keep children safe.

70 UK Gambling Commission, "Age and ID Verification," https://www.gamblingcommission.gov.uk/public-and-players/guide/age-and-id-verification.

71 American Gaming Association, "Responsible Gaming Regulations and Statutes Guide," https://www.americangaming.org/resources/responsible-gaming-regulations-and-statutes-guide/.

72 Gambling Commission, "Complaints and Disputes: Procedural Information, Provision, and Reporting," https://www.gamblingcommission.gov.uk/print/complaints-and-disputes-procedural-information-provision-and-reporting.

Established Gambling Recommendations	Potential GenAI Application	Relevant Safety-by-Design Principles
Auditing and oversight Regular audits of gambling devices are conducted to ensure fairness and compliance. *Ex: The Nevada Gaming Commission, has comprehensive standards for the approval and oversight of gaming devices to ensure their integrity.*[73]	**Contingent on the framework for psychological assessment and independent review, as recommended above, independent audits of GenAI systems could verify that its platform does not intentionally or unintentionally foster dependency or maladaptive behaviors. Furthermore, platforms could provide visibility into their auditing mechanisms, including practices for surfacing unintended consequences.**	→ User safety approached as a shared responsibility; users are not left to manage their own safety.
Vulnerability screening *Ex: The UK Gambling Commission encourages gambling providers to request information about users' income and significant financial changes, such as property purchases, as part of affordability checks to ensure the legitimacy of funds and prevent users from gambling beyond their means.*[74]	**GenAI platforms could ask users if they would like to disclose information about emotional well-being in order to better consider what would be helpful for a user, such as restricting access to features that might exacerbate vulnerabilities.** **While this could provide safeguards, it also raises significant ethical concerns about privacy, consent, and the potential misuse of sensitive health data. See Establish Data Trusts and Interdisciplinary Collaboration below for details on how this may be better operationalized.**	→ User safety is approached as a shared responsibility; users are not left to manage their own safety.

This reference to the regulatory models of the gambling industry is primarily intended to illustrate that certain psychological mechanisms, like those underlying addictive behaviors, can potentially be regulated and are not merely abstract concepts. However, it is important to recognize the limitations and differences of this model, as well as rigorously evaluate its utility, rather than view it as a direct fit for regulating GenAI systems, particularly relational chatbots.

Firstly, it is critical to note that the core risks lie in the very use of irregular rewards and gamification within tools designed for emotional companionship. Unlike gambling, where

73 Nevada Gaming Control Board, "Regulation 14: Manufacturing, Distribution, and Technical Standards," https://gaming.nv.gov/uploadedFiles/gamingnvgov/content/Home/Features/Regulation14.pdf.

74 UK Gambling Commission, "Age and ID Verification."

adults typically enter spaces aware of the high-risk potential of the activity, relational GenAI systems are often presented as tools for connection, support, and well-being, making it harder for users to recognize the ramifications of such design elements, especially when systems are marketed toward vulnerable populations such as children, youth, and individuals experiencing loneliness. Unlike environments where risk is explicit and age restrictions clear, relational chatbots can quietly foster dependency, exert emotional influence, and erode autonomy. In fact, just as gambling has strict age restrictions, and jurisdictions such as Australia are exploring or implementing age gating for social media, many working group members strongly believe that children and youth should not have access to relational chatbots employing addictive design features, such as irregular rewards, due to the significant risks these systems pose to their emotional and psychological well-being along with cognitive development.

Moreover, relying on frameworks that focus on identifying vulnerable individuals risks disproportionately shifting responsibility onto users rather than holding companies accountable for deploying manipulative, unsafe, and potentially harmful systems. It is imperative that responsibility for mitigating these harms does not rest solely with users or hinge on identifying who is most susceptible. The gambling model attempts to achieve this by implementing safeguards and regulatory frameworks aimed at mitigating harm and holding operators accountable. However, applying such a model to relational GenAI systems must go further, ensuring that accountability mechanisms address the unique risks posed by these tools, particularly their ability to exploit emotional vulnerabilities and target younger audiences.

Re-examine FDA regulations for GenAI applications

The inherent plasticity of GenAI systems allows them to be dynamically reconfigured and applied to conversational domains that extend well beyond their creators' initial purposes. As discussed in part 1 of this report, because the FDA bases its jurisdiction on a product's intended use, social GenAI platforms that do not claim to be medical devices—but may be used in the mental health context—operate in a grey area and can effectively sidestep regulatory oversight. This is particularly alarming given preliminary research and anecdotal evidence which indicate that while some apps are not explicitly designed for mental health purposes, GenAI enables users to repurpose these apps in ways that may pose mental health risks.[75]

75 De Freitas and Cohen, "The Health Risks of Generative AI-Based Wellness Apps," 1269–75.

To address the potential misuse of technology in healthcare contexts, even when healthcare is not its primary application, regulatory oversight from bodies such as the FDA and their international counterparts may need to evolve. Further research is needed to explore how these regulatory frameworks might adapt to address these challenges effectively. Promising research directions and potential interventions may include the following (much of which has strong complementary overlap with safety-by-design principles, elaborated further below):

> **Item #8: Reevaluating the FDA's definition of "intended use."** Further investigation could include how the FDA and similar regulatory bodies might expand their definition of "intended use" to include secondary or user-driven applications of GenAI chatbots. This could involve studying approaches for assessing the likelihood that applications can be repurposed to determine necessary disclosures and precautions, for example, via a risk-based tiered system. Such an approach would need to determine realistic standards for reasonable best efforts, appreciating the limitations of preventing or mitigating harm arising from unintended uses. For example, aspirin can be misused, frying pans can be wielded.

> **Item #9: Developing a risk-based tiered system for GenAI applications**. A risk-based tiered system could classify and regulate GenAI applications based on their potential use in safety-regulated contexts—and therefore cause harm—regardless of their intended use. This framework would help assess unintended risks and establish corresponding safeguards for GenAI tools while refining criteria for classification and intervention thresholds (again, realistic standards for obligations would need to be determined). Such benchmarks are also crucial for advocacy groups, auditors, and civil society to assess risks and ensure platforms meet ethical safety standards.

The following outlines a preliminary framework for such a tiered system, offering a starting point for researchers to explore, refine, and adapt in future studies and for potential application by consumer advocacy groups. Further research could study how to require developers of GenAI-powered wellness tools to evaluate edge cases impacting health outcomes, using methods like scenario-based testing to anticipate misuse and design safeguards.

Table 3: Notional Risk-Based Tiered System for GenAI Applications

Tier 4
Critical Risk

Tools explicitly designed for diagnostic or therapeutic purposes, falling under medical device regulation. These tools directly address mental health, either as a primary function or through specialized interaction capabilities.

Examples

- AI tools used to diagnose conditions like anxiety or depression.
- Therapeutic chatbots designed to replace or markedly supplement mental health professionals.

Potential risks, including risks from unintended use

- Misinformation leading to misdiagnosis or inappropriate treatment.
- Over reliance on AI without proper professional oversight.
- Emotional dependency leading to maladaptive behavior.

Regulatory and/or best practice measures

- Full compliance with FDA or equivalent medical device regulations.
- Safeguards (for example, as outlined in Table 2: Potential Applications of Gambling Regulations to AI) such as:
  - Defaults or, potentially, hard limits on frequency, engagement, and timing of interactions
  - Transparency and Advertising restrictions, and warnings
  - Real-time warnings during select interactions with users
  - Age restrictions
  - Self-exclusion, "cooling off," and user feedback options
  - Vulnerability screening
  - Behavioral monitoring and intervention planning
  - Auditing and oversight
  - Mandatory partnerships with licensed healthcare providers

Tier 3
High Risk

Tools that engage in emotionally charged and ongoing conversations with users in ways that could mimic therapeutic contexts, even unintentionally, or tools widely repurposed for mental health conversations.

Examples

- Chatbots providing guidance or advice on mental health or well-being.
- Social GenAI companions designed for extensive emotional support or crisis discussions.

Potential risks, including risks from unintended use

- Users treating the tool as a replacement for professional mental health support.
- Emotional dependency leading to maladaptive behavior.
- Escalation of harm due to misinformation or unregulated guidance.

Regulatory and/or best practice measures

- Psychological risk assessment, including user interaction testing with input from mental health professionals.
- Safeguards (for example, as outlined in Table 2: Potential Applications of Gambling Regulations to AI) such as:
  - Defaults or, potentially, hard limits on frequency, engagement, and timing of interactions
  - Transparency and Advertising restrictions and warnings
  - Age restrictions
  - Self-exclusion, "Cooling Off," and user feedback options
  - Vulnerability screening
  - Behavioral monitoring and intervention planning
  - Auditing and oversight

Tier 2
Moderate Risk

Tools designed for narrow wellness applications, productivity, or skill-building, which may be repurposed for mental health interactions but lack explicit therapeutic intent.

Examples

- Virtual assistants for time management or mindfulness reminders.
- Chatbots for journaling prompts.
- GenAI platforms that allow users to practice public speaking or language learning.

Potential risks, including risks from unintended use

- Users inappropriately seeking mental health advice from the tool.
- Emotional dependency or misinformation.

Regulatory and/or best practice measures

- Basic risk assessment and documentation of safeguards.
- Mandatory inclusion of disclaimers about non-therapeutic use.
- Built-in features to redirect users discussing sensitive topics to professional resources, including crisis hotlines.
- Regular reviews to assess patterns of misuse or unintended consequences.

Tier 1
Low Risk

GenAI applications that operate within clear, narrow boundaries aligned with their intended use, focusing on non-sensitive domains.

Examples

» Utility apps like grammar or spelling checkers.
» FAQ chatbots with predefined, non-sensitive responses.

Potential risks, including risks from unintended use

» Misinterpretation of functionality (e.g., users expecting more than the tool offers).

Regulatory and/or best practice measures

» Clear disclaimers stating the tool's limited scope.
» No extensive oversight required, but transparency in data use is mandated.

Shift Internal Organizational Behavior

"Internal and external governance functions can really only intervene at, 'Well, Don't do this, don't do that,' but they don't weigh in quite as much on the 'Well, where do we drive the technology?' So early stage conversations that try to expand people's moral imagination, in particular with chatbots, can help break down intractable claims that people have about things like empathy, help them get more conceptually specific, and also challenge their intuitions about what is good, what good looks like, what their role in producing that might be. And then really help them see lots of gray areas. It's not just a world with chat bots, perhaps it's a world without chat bots, but what are the specific elements that we need to be making choices about in order to drive toward more sort of desirable futures?"

- GII Working Group Member

As discussed above in our section, Modernize Public Policy, regulations nearly always lag behind technological advancements. Traditional "hard controls" (e.g., policy mandates, compliance protocols, review boards) are vital guardrails but may come into play too late in the product life cycle—typically during the evaluation of nearly finalized designs or as reactionary measures after ethical concerns arise. In many organizations, "soft controls" (e.g., values, assumptions) may influence behavior, but are often deeply ingrained and operate

below the surface, making them difficult to assess or modify.[76] It's important to note that there's no one-size-fits-all method for meeting these challenges. Additionally, any approach requires adaptation for particular environments. Factors such as organizational structure, product lifecycle processes, Trust & Safety team mandates, and company culture are all determinative. The following internal organizational behavior mechanisms are one approach to addressing this policy vacuum, suggesting ways that organizations can proactively embed ethical considerations into their development processes.

A bottom-up ethical approach

The norms and values within technology teams influence the technologies they create. While existing approaches like governance, regulation, and ethics training are essential, they do not fully address the day-to-day ethical decision-making dynamics within autonomous, team-driven structures typical of many of today's tech companies. To address this, promising interventions and research directions may include exploring bottom-up ethical approaches, such as the Moral Imagination exercise, which are aimed at fostering a pervasive culture of responsible innovation at the early developmental stages of tech organizations. This methodology was experimented in over 50 workshops at Google.[77]

Within the Moral Imagination framework, the focus is on engineering teams as a whole, which are often cross-functional and involve collaboration among individuals from diverse technical and non-technical backgrounds. These teams often operate with significant autonomy, particularly in the early stages of product and research development, where critical decisions about which problems to address and how to solve them are made before elevating recommendations to higher management. (Depending on the company, it may be more applicable to incorporate additional teams, such as product management, into this process as well.) This autonomy allows teams to have substantial influence over the ethical and societal impact of the technology they create. However, engineering norms that prioritize technical performance, scalability, and speed-to-market may inadvertently deprioritize considerations of societal impact or inclusivity.[78] In the case of GenAI, such norms can lead to systems that perpetuate harmful dynamics, such as amplifying divisive narratives or fostering addictive usage patterns. By focusing on engineers, the Moral Imagination methodology seeks to embed ethical practices directly into the workflows and decision-making processes of the groups most responsible for shaping technology, ensuring ethical considerations are

76 Benjamin Lange, Geoff Keeling, Amanda McCroskery, Ben Zevenbergen, Sandra Blascovich, Kyle Pedersen, Alison Lentz, and Blaise Agüera y Arcas, "Engaging Engineering Teams Through Moral Imagination: A Bottom-Up Approach for Responsible Innovation and Ethical Culture Change in Technology Companies," *Google Research*, October 31, 2023, https://arxiv.org/abs/2306.06901.

77 Lange et al., "Engaging Engineering Teams Through Moral Imagination."

78 Lange et al., "Engaging Engineering Teams Through Moral Imagination."

integrated from the ground up and creating a cascading effect throughout the product life cycle.

In order to demonstrate how researchers and industry practitioners may use the Moral Imagination exercise as the basis for a bottom-up ethical approach to explore, refine, and implement effectively within organizational contexts, we outline its basic structure:[79]

STEP 1: REFLECTION

The first step focuses on helping teams externalize their existing moral intuitions, beliefs, and values related to their work. Through semi-structured discussions, team members articulate their personal motivations, the perceived benefits of their technology, and their vision for a world in which it is fully deployed. Teams evaluate whether their current plans align with these values, identifying any potential trade-offs or tensions. This step aims to make implicit norms explicit, enabling critical evaluation and adjustment as needed, while also clarifying the team's ethical starting point and establishing a shared vocabulary for further dialogue.

STEP 2: EXPANSION

The second step broadens a team's ethical awareness by challenging their perspectives and revealing gaps in their understanding. Teams engage with techno-moral scenarios set five to ten years into the future, which highlight potential societal and ethical challenges their technology may pose. Role-playing exercises encourage participants to adopt diverse stakeholder viewpoints, uncovering new value tensions and ethical considerations. Inclusion exercises, such as the "veil of ignorance," help identify overlooked stakeholder groups and their needs. Additionally, brainstorming on socio-technical harms enables teams to anticipate and mitigate possible adverse impacts. This phase expands ethical perspectives, encourages critical and creative thinking, and ensures stakeholder needs and societal implications are thoroughly explored.

STEP 3: EVALUATION

The final step equips teams to systematically address ethical dilemmas and navigate value trade-offs. Facilitators introduce ethical reasoning tools and frameworks, such as weighing competing moral values, understanding trade-offs, and applying different ethical paradigms like deontology or consequentialism. This step empowers teams to effectively navigate ethical gray areas, providing them with a pluralistic and rigorous foundation for ongoing ethical deliberation. It fosters a structured approach to resolving dilemmas and builds the skills necessary for navigating complex moral landscapes in their projects.

79 Lange et al., "Engaging Engineering Teams Through Moral Imagination."

STEP 4: ACTION

The fourth, and arguably most important, step of the Moral Imagination methodology, Action, focuses on translating ethical insights gained during the workshops into concrete team practices. This is elaborated below.

Align employee incentives

As discussed, organizations may prioritize long-term, high-volume use cases for a GenAI conversational agent to ensure continued engagement, and thus, continued revenue and brand awareness. Employees may have incentives such as bonuses, promotions, or recognition tied to defined objectives and key results (OKRs) that reflect such goals. In fact, excessive focus on incentives or OKRs may lead to "gaming the system," where employees optimize for metrics and de-emphasize work that falls outside the scope of their OKRs, even if it is ethically important.[80] Thus, organizations must ensure that employees' rewards are not tied to proxies for engagement, like longer screen time, that may encourage them to chase behaviors that are inadvertently misaligned.

After completing step four of the Moral Imagination methodology, the responsibility objectives are then formalized into team workflows, including team-level OKRs and product requirement documents (PRDs), ensuring that ethical considerations move beyond theory and are integrated into practical processes and outputs. This approach establishes accountability for ethical commitments by embedding them in structured, actionable frameworks that guide the team's ongoing work. By embedding these commitments into measurable and visible structures, this approach not only establishes accountability but also helps shift team incentives. Aligning ethical objectives with key performance indicators ties ethical behavior to team success, encouraging individuals to prioritize responsible innovation as a central component of their work and making ethical practices a tangible, rewarded aspect of organizational performance.

The Moral Imagination framework is just one example of a bottom-up ethical approach. Other frameworks, such as the Ethical Operating System, employ similar non-didactic and participatory methodologies and share the goal of fostering ethical outcomes within organizations.[81] These bottom-up approaches provide a valuable foundation for researchers and industry practitioners to explore, refine, and implement effectively within organizational contexts. Again, it is important to note that there are no panaceas, and any methodology

80 Steven Kerr, "On the Folly of Rewarding A, While Hoping for B," in *Leadership: Understanding the Dynamics of Power and Influence in Organizations*, ed. R. P. Vecchio, 2nd ed. (Notre Dame, IN: University of Notre Dame Press, 2007), 228–238, https://doi.org/10.2307/j.ctvpg85tk.23.

81 Paula Goldman and Raina Kumra, "Introducing the World's First Ethical Operating System," *Omidyar Network* (blog), August 7, 2018, https://medium.com/omidyar-network/introducing-the-worlds-first-ethical-operating-system-7acc4abc2bfa.

should be adapted to fit the environment. Further promising research and interventions may include:

> **Item #10: Organizational studies and empirical analysis of implementation**. Study the outcomes of applying such bottom-up ethical approaches within technology companies. Focus on identifying patterns and areas for improvement, particularly during goal-setting processes like OKRs.

> **Item #11: Exploring a diverse set of bottom-up ethical approaches**. Investigate how ethical frameworks, such as the Moral Imagination exercise, can be developed and scaled in a way that integrates seamlessly with the autonomous and entrepreneurial culture of technology teams. Emphasize alignment with existing workflows and decision-making structures to foster adoption and impact.

> **Item #12: Adapting bottom-up ethical approaches to organizational contexts**. Examine how bottom-up ethical frameworks can be effectively adapted to different organizational sizes and hierarchies. Investigate strategies to ensure these approaches remain relevant and impactful across diverse structures, from agile startups to complex corporate environments.

Explore Technical Interventions & Approaches to Alignment

> *"I do think that we have ways of talking about ethics that shouldn't just become so confused that we can never agree on anything, because otherwise none of these projects matter. I don't want to end today with, well, we still wouldn't know what to do, because all norms are relative."*
>
> - GII Working Group Member

Technically introducing value judgments and norms in GenAI systems to foster social cohesion raises critical questions: who holds the authority to make these decisions and on what basis? As working group members consistently discussed, this approach runs the risk of prioritizing conformity, inadvertently imposing cultural homogenization and exercising undue paternalism. However, there are technical interventions that preserve autonomy while also prioritizing safety and paradigms of alignment that approach the task with greater sensitivity, inclusivity, and an awareness of these risks. The following approaches to technical interventions and alignment build on this notion and serve as a roadmap for researchers and developers to

consider the mechanics of GenAI systems and platforms from the perspective of social cohesion.

Adopt a shared decision-making model

"This brings up a very fundamental question here: do I, as the user of this thing, have the right to say, 'I don't want any conflicting input.' And it would be hard for someone to say, 'No, you don't have that right. You must be exposed to conflicting input.' (...) The more basic value point is, should we allow systems that allow you to only get input that you want?"

"But it has to do with what the companies are allowed to create, not the rights, individual rights."
- GII Working Group Members

A fundamental conflict exists between individuals' desire to control their personal informational landscapes and society's normative goal of promoting diverse perspectives, values, and attitudes. However, as with any industry, the development and deployment of technologies are impacted by prevailing economic incentives. These incentives often address users' immediate preferences and desires, as indicated by their revealed preferences (i.e., observable choices like clicks, engagement time, etc.), making them easy to collect and appealing to users. However, there is growing recognition that revealed preferences and normative/informed preferences (i.e., the choice a person would make under ideal conditions) widely differ. It is important to clarify that this distinction does not imply an external authority can determine what is "best" for an individual but rather highlights systemic factors that shape and sometimes constrain the range of available choices, that a user may have rather chosen such had been known. This is due to:[82]

- **Passive choice**, which occurs when individuals accept default options, often out of procrastination or inattention, leading to suboptimal outcomes. This is often a result of complexity in decision-making which can cause individuals to avoid decisions, rely on simplistic heuristics, or choose poorly due to misunderstanding.
- **Limited personal experience** prevents individuals from learning what is in their best interest, as they lack sufficient feedback or relevant examples to guide their choices.
- **Third-party marketing** further distorts preferences by leveraging advertising or branding to influence decisions that do not align with individuals' true needs or values.
- Finally, **intertemporal choice**—decisions involving future consequences—often leads to "present bias," where individuals prioritize small but immediate rewards over greater long-term benefits.

82 John Beshears, James J. Choi, David Laibson, and Brigitte C. Madrian, "How Are Preferences Revealed?" *Journal of Public Economics* 92, no. 8–9 (August 2008): 1787–94, https://doi.org/10.1016/j.jpubeco.2008.04.010.

Thus to be genuinely helpful to user well-being, social GenAI must comprehend users' immediate and long-term goals as well as their revealed and informed preferences. Typically, to address issues of this nature, public platforms have implemented ways of sharing administrative responsibilities to its user base. For example, social media platforms typically include user-driven reporting systems to flag inappropriate or harmful content. These approaches distribute governance and decision-making, empowering users to shape the platform's environment in alignment with shared values and norms. However, the 1:1 and sensitive nature of social GenAI interactions makes user-driven governance impractical at scale. This calls for exploring alternative frameworks that prioritize individual agency while addressing the unique challenges of personalized interactions.

One such approach worth considering is the shared decision-making model (SDM) rooted in clinical medical practice. SDM involves clinicians and patients collaboratively making decisions based on available evidence and the patients' preferences and values. Research suggests that such an approach enhances patient knowledge, confidence, and engagement.[83] In the context of GenAI, an SDM could integrate user feedback, promote transparency in the system's recommendations, and allow for dynamic adjustments as user goals evolve over time. By embedding these principles, the technology would foster a partnership model where individual informed preferences are prioritized without presuming to define a user's "true" needs, thereby respecting autonomy and avoiding paternalism. Table 4 depicts a simplified three step clinical SDM approach and how it might be adapted to the social GenAI context.[84]

83 Stacey et al., "Decision Aids for People Facing Health Treatment or Screening Decisions," *The Cochrane Database of Systematic Reviews* 1, no. 1 (2024): CD001431, https://doi.org/10.1002/14651858.CD001431.pub6.

84 Glyn Elwyn et al., "Shared Decision Making: A Model for Clinical Practice," *Journal of General Internal Medicine* 27, no. 10 (2012): 1361–67, https://doi.org/10.1007/s11606-012-2077-6.

Table 4: Three-step clinical SDM approach adapted for the Social GenAI context

Traditional Medical Context	Potential Social GenAI Context
1. Choice Talk	
Introduce the existence of multiple treatment options and involve the patient in the decision-making process.	The bot could start by presenting a range of possible goals or interaction types (e.g., productivity support, mental health assistance, or making new connections) and invite the user to choose or co-create their focus for the duration of usage.
2. Option Talk	
Provide detailed information about the options, including their risks and benefits, and facilitate patient understanding using tools like decision aids.	The bot would then offer tailored options for how it could assist, explaining potential outcomes, trade-offs, and any limitations of its capabilities. Visual aids, conversational prompts, or interactive tools could enhance the user's understanding of their choices and the potential risks associated with them.
3. Decision Talk	
Guide the patient in exploring preferences and making informed decisions, ensuring support throughout the process.	Finally, the bot would work with the user to align on a course of action, iteratively refining its support based on ongoing feedback and ensuring the user feels empowered and supported throughout the interaction.

Such a model could serve as an implicit framework for social GenAI development. Alternatively, a GenAI bot could explicitly adopt an SDM approach at the start of its interaction with a user to foster meaningful, user-centric collaboration. Aligned with the HHH framework, this approach enables the platform to be transparent about its limitations while effectively eliciting a user's normative preferences. This dual focus enhances the platform's ability to provide meaningful support by allowing the user to more meaningfully co-create their experience with the social GenAI platform.

Adopting this model could enhance the alignment of GenAI systems with users' long-term well-being while establishing trust and fostering a more participatory dynamic, bridging the gap between user agency and AI assistance. As such, further research and interventions could be:

Item #13: Defining an appropriate shared decision-making model for GenAI. This involves exploring how such a model can balance the immediate preferences and long-term well-being of users. Special attention should be given to whether the model needs to adapt to specific demographic characteristics, such as age, cultural background, or cognitive capacity, to ensure inclusivity and relevance across diverse user groups.

Item #14: Developing improved methods to meaningfully elicit user preferences. Research should focus on creating tools and frameworks that enable AI systems to understand both revealed and informed preferences accurately. Additionally, determining which proxies—such as behavioral data, explicit user feedback, or contextual factors—are most effective in capturing preferences will be crucial for ensuring the AI's alignment with user goals.

Item #15: Piloting the shared decision-making approach in real-world contexts. Implementing and testing the model in practical settings will provide valuable insights into its feasibility and effectiveness. This includes gathering data on user engagement, adoption rates, and overall satisfaction, as well as identifying barriers to implementation and opportunities for refinement. These insights will help optimize the model for broader application.

Item #16: Flourishing-by-design. Relying solely on revealed preferences risks also ignores GenAI's potential to realize a more profound and beneficial vision of AI—one that actively supports individuals in their personal development and overall human flourishing.[85] While this project centers on preventing harm, participants expressed cautious optimism that thoughtfully designed social GenAI can achieve meaningful, constructive outcomes. Existing research on healthy digital interactions offers a starting point, but must be adapted to social GenAI's unique one-on-one user-tailored context. Further study should refine flourishing frameworks in this environment. Drawing on therapy—where professionals balance empathy, enabling client agency, and boundaries to prevent dependency—could inform GenAI designs that foster autonomy and growth. Such approaches might include gradual detachment features and in-platform group sessions with other users—ideas drawn from client-empowerment and group therapy. Emphasizing flourishing-by-design thus shifts the focus from mere harm reduction to positive growth, human connection, and societal cohesion.

85 Joel Lehman, "Machine Love," *AI Objectives Institute*, February 22, 2023, https://arxiv.org/abs/2302.09248.

Develop AI with stronger metacognitive abilities

This is not to be confused with the metacognitive challenges that GenAI poses (as detailed in part 1 of this report).

Technically aligning AI with human values and the public interest is a deeply complex challenge. These values are often shaped by cultural contexts, can unintentionally perpetuate undesirable norms, and may overlook or marginalize minority perspectives. This is why some scholars, and some of our working group members, argue that such a rigid application of "alignment" is simply not practical at scale. Moreover, many social GenAI platforms will not be used in "closed tasks" or otherwise narrow interactions. As noted, these platforms are often used by humans in open-ended, context-dependent, and nuanced conversations, which can escape the parameters the model is trained to consider appropriately.

Therefore, early researchers propose a shift towards AI systems with advanced metacognitive reasoning capabilities—such as self-reflection, acknowledgment of uncertainty, and adaptive decision-making—thus enabling them to navigate complex scenarios characterized by incommensurable goals, uncertainty, and nonlinear dynamics, much like human cognition is designed to handle.[86] In the context of social GenAI, this would enable AI systems to engage more effectively in open-ended, nuanced interactions by understanding diverse perspectives, adapting to evolving social norms, and balancing conflicting values, thereby addressing the complexities of real-world conversations and interactions. This is a promising avenue of research and technical intervention that involves a subsequent shift in how benchmarking and evaluations are conducted in most labs. Other researchers have highlighted similar pathways for advancing this work, which are briefly outlined and applied to a social GenAI context.[87] As such, researchers and developers may choose to further early research via the following:

> **Item #17: Moving Beyond Outcome-Based Metrics**. Social GenAI chatbots often engage in dynamic, context-dependent conversations, which require reasoning that extends beyond producing accurate or coherent responses. Future research should address the need for chatbots to demonstrate authentic metacognitive reasoning by:
>
> - Developing benchmarks that evaluate how chatbots manage conversational uncertainty, such as recognizing when they lack sufficient knowledge or when a user's question requires clarification or deeper reflection.

86 Samuel G. B. Johnson, Amir-Hossein Karimi, Yoshua Bengio, Nick Chater, Tobias Gerstenberg, Kate Larson, Sydney Levine, Melanie Mitchell, Iyad Rahwan, Bernhard Schölkopf, and Igor Grossmann, "Imagining and Building Wise Machines: The Centrality of AI Metacognition," *arXiv*, October 2024, https://arxiv.org/abs/2411.02478.

87 Johnson et al., "Imagining and Building Wise Machines."

- Creating evaluation frameworks that measure a chatbot's ability to navigate conflicting conversational goals, such as balancing empathetic support with objective truth-telling or reconciling diverse user perspectives.
- Incorporating metrics for explainability in dialogue, assessing whether chatbots can articulate the reasoning behind their responses in a way that fosters transparency with users.

Item #18: Social, Collaborative, and Context-Rich Training. Social GenAI chatbots must navigate conversations that reflect human complexity, including emotional nuances, cultural diversity, and ethical dilemmas. Researchers could explore:

- Designing context-rich conversational datasets that include examples of multi-faceted human interactions, such as conflict resolution, cross-cultural dialogue, or emotionally charged discussions.
- Developing multi-user conversational simulations, where chatbots interact with multiple participants simultaneously, practicing skills like perspective-taking, mediating disputes, and adapting to varying conversational tones.
- Investigating human-in-the-loop training for chatbots, where user feedback dynamically guides the system to improve its ability to identify and adapt to emotional cues, cultural sensitivities, and evolving conversational contexts.

Item #19: Balancing Generalization and Specialization. Social GenAI needs to excel in specific domains (e.g., mental health support, customer service) while maintaining the flexibility to generalize across diverse conversational scenarios. To achieve this, research could focus on:

- Designing conversational scenarios that emphasize cross-domain adaptability through metaphorical reasoning, such as helping users draw lessons from one social domain (e.g., conflict resolution in workplace settings) to apply to another (e.g., managing familial disputes).
- Developing mechanisms to enable chatbots to transfer conversational strategies (e.g., empathy, de-escalation techniques) learned in one domain to interactions in another domain.
- Investigating how chatbots can leverage external expertise dynamically, such as integrating responses from domain-specific knowledge bases while maintaining conversational fluency and emotional intelligence.

Introduce new Cognitive Forcing Functions

> *"And similarly, in another domain, we built a fully explainable model for hiring. Literally, not a single one of our customers ever looked at the explanations. So I'm simply throwing this out that, yeah, you can build stuff like this. It isn't science fiction. But then the human factor comes in. Is anybody going to use it?"*
>
> - GII Working Group Members

People frequently overtrust AI's suggestions, even when they are incorrect, leading to suboptimal decisions. According to dual-process theory, human decision-making often relies on fast, emotional, heuristic-based "System 1" thinking, which can lead to metacognitive flaws such as anthropomorphizing the chatbot, ascribing it agency, or not having the capacity to critically think about its outputs. "System 2" thinking, which is slower and more analytical, can mitigate these errors but requires effort - which the DCDI and GII work has shown, humans are cognitively disinclined to do.[88] Explainable AI systems, which provide rationales and confidence levels for AI decisions, were initially believed to engage System 2, encourage more critical thinking, and thus prevent such overreliance. However, early studies demonstrate this may not be the case. In fact, sometimes, the presence of explanations increases trust and overreliance, as they are taken for indicators for competence.[89]

Thus, it may be the case that the best efforts to be transparent and explain a social GenAI's limitations may fall short of its intended purpose. Thus, researchers have proposed Cognitive Forcing Functions (CFFs) as an intervention to disrupt System 1 thinking and encourage users to engage System 2 thinking when interacting with AI.[90] By introducing deliberate friction or requiring additional cognitive effort during decision-making, CFFs aim to reduce blind trust in AI outputs and promote more critical evaluation of its suggestions. Researchers tested three CFFs in their study:[91]

- **"On Demand"** required participants to explicitly request AI suggestions by clicking a button, aiming to foster critical thinking by making users actively seek the AI's input.
- **"Update"** had participants make an initial decision without AI assistance and then presented the AI's suggestion and explanation, allowing them to revise their choice. This design was intended to encourage users to compare their judgment with the AI's input critically.

88 Daniel Kahneman, *Thinking, Fast and Slow* (New York: Farrar, Straus and Giroux, 2011).

89 Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld, "Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21 (New York, NY: Association for Computing Machinery, 2021), 1–16, https://doi.org/10.1145/3411764.3445717.

90 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, no. CSCW1 (April 22, 2021): 188:1–188:21, https://doi.org/10.1145/3449287.

91 Buçinca, Malaya, and Gajos, "To Trust or to Think."

» **"Wait"** introduced a mandatory 30-second delay before revealing the AI's suggestion, during which participants were encouraged to form their own hypothesis about the task. This delay sought to compel users to engage in preliminary analytical thinking before being exposed to AI recommendations. Together, these interventions were designed to shift users away from heuristic-based decision-making and toward more deliberate evaluation of AI-generated outputs.

The study found that CFFs significantly reduced overreliance on AI compared to simple explainable AI approaches. When AI predictions were incorrect, participants in CFF conditions were more likely to spot and disregard the flawed suggestions, resulting in better decision-making outcomes. However, even with CFFs, overreliance was reduced but not entirely eliminated.[92]

Although CFFs are by no means an all-encompassing solution to social GenAI's negative effects, they do provide a supplementary technical intervention that may help mitigate overreliance on AI by encouraging more deliberate and critical engagement with AI-generated outputs. Technical solutions such as these, much like screen time reminders, are suggestive nudges but can be bypassed or otherwise ignored. As such, researchers and developers can use these preliminary CFFs studies as the foundation for other cognitive interventions that are both effective and accessible to a diverse range of users:

> **Item #20: More research into the efficacy, development, and uptake of CFFs**. The current study provides a starting point for understanding how CFFs can reduce overreliance on AI, but there is a need to explore additional types of cognitive interventions tailored to different decision-making contexts, including social GenAI interfaces and critical decision making contexts. However, examining the uptake of such interventions in tech companies is crucial, as their implementation might face resistance due to concerns that these measures could slow down user experiences—again, highlighting the tension between the principles of helpfulness and harmlessness.

> **Item #21: Further research in human-computer interaction**. Such research is needed to address the notable trade-offs observed in this study. While CFFs were effective at reducing overreliance on AI, they were also perceived as more mentally demanding and complex, resulting in lower trust and user preference compared to simpler Explainable AI interfaces. These findings underscore the challenges of designing AI systems that adhere to the principles of helpfulness, harmlessness, and honesty. Specifically, improving honesty and reducing overreliance (as achieved through CFFs) can negatively impact usability and equitable accessibility, potentially diminishing the system's perceived helpfulness and

92 Buçinca, Malaya, and Gajos, "To Trust or to Think."

inclusivity. Addressing these trade-offs is crucial for developing AI systems that balance effectiveness with user acceptance and equity.

Harness insights from affective computing

Affective computing research has shown that emotions often considered "negative," such as confusion and frustration, are healthy and normal indicators of the learning process.[93] Such emotions are often positively co-opted by expert teachers, and can also be positively co-opted by technology to encourage increased engagement, understanding, and retention.[94] Insights from affective computing, particularly textual analysis, could enable social GenAI to function as a more effective mediator by dynamically responding to user outputs that reflect specific emotional states. This is an effective precursor to many of the interventions cited in this report; by offering more precise signaling, such technology could preemptively identify when human oversight is required, recommend the allocation of additional resources, or recognize instances where users' informed preferences are not being adequately met.

However, as discussed in Part 1, GenAIs can have the tendency to perpetuate prescriptive ideas of the "correct" way to speak, think, and feel. Affective computing risks exacerbating this problem if it evaluates user data against a monolithic or universalized notion of "appropriate" emotional expression.[95] To avoid such pitfalls, research in this domain must adopt a contextualized and inclusive approach, recognizing the diverse and fluid ways individuals communicate and express emotions across cultural and situational boundaries. It should be responsive to the ways vernacular and cultural norms evolve over time, rather than reinforce rigid, normative standards of emotional or linguistic expression. Furthermore, there are significant privacy concerns regarding how this highly sensitive emotional and behavioral data is collected, analyzed, and stored. Safeguards must be put in place to ensure transparency, user consent, and the ethical use of such data, as misuse could exacerbate surveillance risks, deepen power imbalances, and lead to unintended harms for marginalized or vulnerable communities. Such privacy concerns could be addressed via developing a data trust, as elaborated later on later in this report.

Building on these considerations, integrating such capabilities could significantly enhance the system's ability to support nuanced human-AI collaboration. This would make interactions not

93 B. Kort, R. Reilly, and R.W. Picard, "An Affective Model of Interplay between Emotions and Learning: Reengineering Educational Pedagogy-Building a Learning Companion," In *Proceedings IEEE International Conference on Advanced Learning Technologies*, (Madison, WI: Proceedings IEEE Xplore, August 2001): 43–46, https://doi.org/10.1109/ICALT.2001.943850.

94 Walter Burleson and Rosalind W. Picard, "Affective Agents: Sustaining Motivation to Learn Through Failure and a State of 'Stuck'," *MIT Media Lab Publications*, 2004, https://www.media.mit.edu/publications/affective-agents-sustaining-motivation-to-learn-through-failure-and-a-state-of-stuck/.

95 Kerry McInerney and Os Keyes, "The Infopolitics of Feeling: How Race and Disability Are Configured in Emotion Recognition Technology," *New Media & Society* (2024): 1–19, https://doi.org/10.1177/14614448241235914.

only safer but also more contextually aware, ultimately allowing GenAI systems to become more helpful and harmless to users. As such, researchers may find it promising to:

Item #22: Developing models for detecting subtle emotional cues. Research should prioritize the development of computational models capable of accurately identifying subtle emotional cues, such as confusion, frustration, engagement, or signs of heightened distress or franticness, through advanced linguistic analysis. These models must incorporate contextual parameters, including syntactic structures, lexical choices, and paralinguistic markers, to enhance their sensitivity and precision. Additionally, they should account for cultural and situational variability in emotional expression, ensuring that the models are adaptable to diverse communication styles and vernaculars. Leveraging the interplay between natural language processing (NLP) and affective computing methodologies would enable more sophisticated recognition of nuanced emotional states, contributing to the overall effectiveness of human-computer interaction. Furthermore, ethical considerations such as privacy, data security, and transparency must guide the development process, ensuring that emotional data is used responsibly and equitably. By balancing technical innovation with inclusivity and ethical safeguards, these models can foster more empathetic, context-aware, and human-centered AI systems.

Item #23: Dynamic adjustment of responses based on emotional cues. Utilizing detected emotional signals, social GenAI systems should dynamically tailor their responses to align with users' affective and cognitive states in real time. For instance, when confusion is detected, the system could reframe its output using clearer, more accessible language or provide illustrative examples to facilitate comprehension. Such adaptability could significantly enhance the contextual relevance and user-centricity of AI-mediated interactions.

Evolve Frameworks and Data Collection Methodologies for Understanding AI-Human Interaction

Investigate a new paradigm of research

Much AI-human interaction research has been conducted within the framework of the "Computers are Social Actors" (CASA) paradigm, which posits that users tend to apply human-human social scripts, patterns of behaviors and responses typically reserved for human-human interaction, when engaging with computer agents that do not necessarily warrant such. However, a key limitation of applying CASA in its original form today is its reliance on assumptions about human-technology relationships that were shaped in the 1990s. At that time, social media had not yet been invented (or was in its infancy) and human interactions with computers were infrequent, task-specific, and often mediated through relatively primitive interfaces.[96]

Therefore, researchers argue that CASA fails to account for the profound changes in technology and society over the past three decades.[97] Technologies now occupy a central and continuous role in daily life, and users have developed more nuanced understandings and expectations of their capabilities. Additionally, GenAI can engage in adaptive, personalized, and emotionally resonant interactions, beyond the capabilities envisioned under the original CASA paradigm. By adhering to these outdated assumptions, CASA may unintentionally reinforce a perspective that oversimplifies AI interactions, failing to fully account for the nuanced ways users understand and interact with these technologies in the modern context.

Researchers propose investigating a revised CASA framework in which humans may develop "human-media social scripts," distinct from human-human social scripts, through repeated

96 Andrew Gambino, Jesse Fox, and Rabindra Ratan, "Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm," *Human-Machine Communication* 1 (February 1, 2020): 71–86, https://doi.org/10.30658/hmc.1.5.

97 Gambino, Fox, and Ratan, "Building a Stronger CASA."

and prolonged interactions with media agents such as GenAI conversational systems.[98] For example, consider prompt engineering, the practice of crafting specific inputs to guide and influence GenAI responses. Unlike the application of human-human script to a computer, where interactions follow socially ingrained norms and expectations, prompt engineering involves an understanding that specific inputs can deliberately shape specific outputs. This intentional dynamic creates a distinct human-media script, as the interaction is guided by the user's awareness of the system's programmable nature. Such an updated CASA framework investigates AI-human interaction as a separate category of relational experience that may fulfill specific social needs in distinct ways. As such, building this framework requires researchers to take a multifaceted approach, including:

> **Item #24: Longitudinal studies of GenAI use**. These are essential for tracking behavioral changes and relational dynamics over extended periods, providing insights into how users adapt to and integrate media agents into their daily lives. Longitudinal approaches enable researchers to observe patterns of trust development, shifts in reliance, and the emotional trajectories of interactions with media agents. They also allow for the identification of when and how human-media interaction scripts begin to diverge from human-human interaction scripts, offering a clearer understanding of the distinct relational frameworks users apply over time. Furthermore, longitudinal studies would allow researchers to investigate how interactions with social GenAI agents shape human relationships and social behaviors over time.

> **Item #25: Studies of more varied GenAI use cases**. Much existing research focuses on extreme or edge cases, such as scenarios where machines are seen as full replacements for human relationships. This narrow focus overlooks the more nuanced and pressing question of the nature of relationships with social GenAI agents when they function as supplementary, rather than replacement, connections.[99] Research should explore contexts where AI companions coexist with and enhance human relationships, examining their role as an additional option within broader social ecosystems. These studies would provide insights

98 Gambino, Fox, and Ratan, "Building a Stronger CASA."

99 Eva Weber-Guskar, "How to Feel about Emotionalized Artificial Intelligence? When Robot Pets, Holograms, and Chatbots Become Affective Partners," *Ethics and Information Technology* 23, no. 4 (December 2021): 601–10, https://doi.org/10.1007/s10676-021-09598-8.

into how media agents complement existing social structures, fulfill distinct needs, and shape the dynamics of interpersonal and human-machine interactions.

Establish "data trusts" and interdisciplinary collaboration

Throughout this research agenda, there are calls for greater research to better understand the context of users' interactions with social GenAI tools, and users' behavior (e.g. frequency, duration, observed norms, etc.). These insights are also paramount in designing effective interventions that can mitigate the effects of dependency, emotional manipulation, and potential exploitation, ensuring that these technologies are used responsibly and ethically while minimizing harm to users. However, collecting such data raises significant privacy concerns, including a loss of anonymity, data misuse, and a lack of informed consent. Furthermore, a lack of trust in such data collection may discourage users from freely engaging with these tools, potentially skewing the very research the data aims to support. Thus, one potential solution is the establishment of a governance mechanism known as a "data trust."[100]

A data trust is a legal framework that gives an independent entity (the trustee) the responsibility to hold and manage data on behalf of a group of beneficiaries (the users).[101] The trustee is meant to act in the best interest of users and can define access, use, and liability to align with agreed-upon ethical guidelines, privacy standards, and value-driven goals. They are often referred to as a solution to create "fiduciary accountability"[102] in situations in which pools of data are of interest to multiple stakeholders.

For example, consider a non-profit entity or third-party coalition established to serve as the trustee for user data generated on a conversational social GenAI platform:

- The trustee body could be composed of experts in diverse fields such as AI ethics, privacy law, mental health, technology, and human flourishing. This interdisciplinary expertise ensures the ability to address complex trade-offs and prioritize outcomes that serve the collective interests of the users.
- Trustees would only grant access to this data under controlled conditions that align with the best interests of its users. Such conditions could include purposes like specific research, targeted interventions, or algorithm improvements aimed at fostering community wellbeing, implementing safeguards, and enhancing understanding of user

100 George Zarkadakis, "Data Trusts Could Be the Key to Better AI," *Harvard Business Review*, November 2020, https://hbr.org/2020/11/data-trusts-could-be-the-key-to-better-ai.

101 Bianca Wylie and Sean Martin McDonald, "What Is a Data Trust?," *CIGI*, October 9, 2018, https://www.cigionline.org/articles/what-data-trust/.

102 Canadian Heritage, "Digital Content Governance and Data Trusts — Diversity of Content in the Digital Age," *Government of Canada*, https://www.canada.ca/en/canadian-heritage/services/diversity-content-digital-age/digital-content-governance-data-trust.html#a4.

behavior to support precise interventions and optimize resource allocation. Furthermore, trustees would retain the authority to revoke access if data recipients fail to adhere to the agreed-upon terms and conditions.[103]

» A key function of a data trust is managing data collection and access, while ensuring users retain ownership and control. Users could opt in or out, request deletion, and review data usage. The trust could also require data to be anonymized and aggregated, reducing misuse risks while enabling valuable behavioral insights

Data trusts, while a valuable tool to progress many of the research interventions recommended in this report, are not a panacea. They can be misused and are only as effective as the trustees and governance mechanisms supporting them. Robust legal frameworks, policy measures, and rights-based approaches are essential to ensure ethical and effective governance. As such, further promising research methods and interventions may include:

Item #26: Investigating data trust in the context of GenAI. Further research is required to investigate applicability of data trusts to GenAI platforms, where issues such as intellectual property rights, privacy, and algorithmic accountability create unique governance hurdles. This includes studying how data trusts can mediate conflicts between stakeholders, ensure the execution of fiduciary duty, and incorporate safeguards against misuse of sensitive data. Additionally, exploring how data trusts can adapt to the fast-evolving nature of GenAI technologies, including managing data from decentralized and collaborative models, is critical.

Item #27: Piloting data trusts in real-world applications. Beyond research, more pilots are needed to implement data trusts and study their practical applications. These pilots can provide valuable insights into operational challenges, such as establishing trust boundaries, integrating with existing legal systems, and creating mechanisms for ongoing accountability and stakeholder feedback. This may be particularly instrumental for the collection of data required for more longitudinal studies.

Conclusion

"Technology proposes itself an architect of our intimacies."

- Professor Sherry Turkle[104]

The working group identified a paradox: GenAI may erode our own metacognitive processes, yet building these AI systems with stronger metacognitive capabilities might help humans

103 Anouk Ruhaak, "Data Trusts: A New Approach to Protection," *The RSA*, June 2020, https://www.thersa.org/blog/2020/06/data-trusts-protection.

104 Jeffry Pike, "Turkle Talks Technology, Intimacy," *Harvard Gazette*, May 13, 2010, https://news.harvard.edu/gazette/story/2010/05/turkle-talks-technology-intimacy/.

retain autonomy. Wisdom, once viewed as a hallmark of human cognition, is now cast as an aspirational quality of AI—enabling it to navigate complexity, uncertainty, and ethical dilemmas. By embedding self-reflection, tolerance of uncertainty, and adaptive decision-making into GenAI, we attempt to encode the very traits we risk losing in ourselves through over-reliance and unthoughtful design. A preference for texting over phone calls did not exist before texting, just as a preference for phone calls over in-person meetings did not exist before the telephone. The rise of generative AI as a social actor and our increasing inclination to interact with it is a still-evolving phenomenon—one that demands careful attention to the trade-offs these changes entail.

The approach outlined in this report urges policymakers and technologists to more practically consider these complexities, focusing on building infrastructures that ensure user safety and autonomy, while at the same time facilitating this technology's potential to foster human flourishing. Central to this effort is aligning GenAI development and deployment with principles like Helpfulness, Honesty, and Harmlessness, addressing challenges in cognition, social trust, and social cohesion. Achieving this balance involves a combination of public policy measures, organizational best practices, technological interventions, and ongoing research. Ultimately, success depends on a unified commitment from policymakers, technologists, and civil society to establish an inclusive, accountable digital civic infrastructure.

> *"Attachment is kind of a misnomer, because it's defined in terms of disruption, not actual attachment. Everyone is happy when they're getting what they want. The difference in attachment emerges when people don't get what they want. (...) But I can't really conceive of an AI that would reject you and decide to not speak to you, which a human could do."*
>
> - GII working group member

INSTITUTE FOR SECURITY AND TECHNOLOGY

www.securityandtechnology.org

info@securityandtechnology.org