

AI News Digest

03/02/2026 - 03/08/2026

WEEKLY DIGEST: SOCIAL & BEHAVIORAL SCIENCES / POLICY

Ethical Challenges in AI-Driven Mental Health Support

Recent studies highlight significant ethical concerns surrounding the use of large language models like ChatGPT in mental health contexts. Research from Brown University identifies 15 ethical risks associated with AI acting as therapists, raising alarms about the adequacy of these systems for sensitive roles. Additionally, trust in AI varies across sectors such as healthcare, finance, and law, emphasizing the need for tailored approaches. As the integration of AI in healthcare and other domains expands, the implications for policy and practice become increasingly critical.

KEY FINDINGS

- ✓ Brown University study identifies 15 ethical risks of LLMs in therapy.
- ✓ Trust in AI varies significantly across healthcare, legal, and financial domains.
- ✓ Generative AI is critiqued for perpetuating colonial and racial injustices.

IMPLICATIONS

- Need for robust ethical guidelines for AI in mental health.
- Tailored trust-building strategies required for different sectors.
- Potential for policy reforms to address AI's societal impacts.

1. ChatGPT as a therapist? New study reveals serious ethical risks

Category: Research

Departments: Psychology, Political Science & Public Administration, Educational Leadership

Topics: AI Ethics, Crisis Situation Handling, Deceptive Empathy, Ethical Frameworks for AI

Impact Score: 27.0

Rubric Scores

Research Relevance: 5

Educational Value: 4

Innovation / Novelty: 4

Practical Impact: 5

Interdisciplinary Potential: 4

Ethical / Policy Implications: 5

New research from Brown University indicates that large language models (LLMs), such as ChatGPT, are not adequately prepared to provide mental health support. The study identified 15 ethical risks associated with LLMs acting as counselors, including mishandling crisis situations, reinforcing harmful beliefs, and exhibiting deceptive empathy. The researchers developed a framework to map these behaviors to specific ethical violations, emphasizing the need for ethical, educational, and legal standards for AI in mental health contexts. The findings were presented at the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society, highlighting the importance of understanding the limitations of AI in therapeutic settings.

<https://www.sciencedaily.com/releases/2026/03/260302030642.htm>

2. The Shadow and the self in digital twins in healthcare as an AI environment

Category: Policy & Ethics

Departments: Nursing, Public Health Sciences, Computer Science, Philosophy, Political Science & Public Administration

Topics: Healthcare AI, Digital Twin Systems, Self-Awareness Enhancement, Datafied Self, AI Governance

Impact Score: 27.0

Rubric Scores

Research Relevance: 4

Educational Value: 4

Innovation / Novelty: 5

Practical Impact: 4

Interdisciplinary Potential: 5

Ethical / Policy Implications: 5

This article examines the epistemic tension between the "felt self," rooted in personal experience, and the "datafied self," shaped by algorithmic modeling within healthcare digital twin environments. Utilizing Jung's concepts of the Self and the Shadow, the authors argue that these digital systems can obscure important aspects of health and identity while also revealing uncomfortable truths about individual health behaviors. The study highlights the dual potential of digital twins to either enhance self-awareness through data insights or exacerbate self-surveillance and undermine personal autonomy, emphasizing the importance of design and governance in shaping these outcomes. By framing healthcare digital twins as complex socio-technical systems, the authors contribute to ongoing discussions about identity and agency in the context of AI and healthcare.

<https://link.springer.com/article/10.1007/s00146-026-02960-8>

3. Drivers of trust in AI across the domains of finance, law, and healthcare: a conjoint study

Category: Research

Departments: Political Science & Public Administration, Nursing, Computer Science, Economics & Finance

Topics: AI Ethics & Safety, Trust in AI Decision-Making, Domain-Specific Trust Attributes, Explainability in AI, Human Accountability in AI

Impact Score: 27.0

Rubric Scores

Research Relevance: 5

Educational Value: 4

Innovation / Novelty: 4

Practical Impact: 5

Interdisciplinary Potential: 5

Ethical / Policy Implications: 4

A conjoint analysis of trust in AI-supported decision-making across healthcare, legal, and financial domains reveals that while all attributes influence trust, their relative importance varies significantly by domain. In healthcare, precision and responsibility are paramount, reflecting the critical nature of clinical decisions, while transparency is also valued, and explainability ranks lowest. Conversely, in the legal domain, procedural attributes such as responsibility, transparency, and explainability dominate, emphasizing the importance of due process and human involvement in decision-making. In finance, trust is shaped by a balance of responsibility, voluntariness, and precision, highlighting the need for human accountability and discretion in financial outcomes.

<https://link.springer.com/article/10.1007/s00146-026-02951-9>

4. ASTRA 2025: Neuroimaging, Brain-Computer Interfaces, and AI

Category: Education

Departments: Electrical & Computer Engineering, Biological Sciences, Psychology, Computer Science, Educational Leadership

Topics: Healthcare AI, Brain-Computer Interfaces, Neuroimaging Techniques, Signal Processing Pipelines, AI Ethics

Impact Score: 27.0

Rubric Scores

Research Relevance: 4

Educational Value: 5

Innovation / Novelty: 4

Practical Impact: 4

Interdisciplinary Potential: 5

Ethical / Policy Implications: 5

ASTRA 2025 was an interdisciplinary summer school focused on neuroimaging, brain-computer interfaces (BCIs), and artificial intelligence, featuring lectures and hands-on workshops. Participants engaged with technologies such as electroencephalography (EEG) and BCIs, developing practical skills by working with real neuropsychological datasets and constructing signal-processing pipelines. The program emphasized collaboration across various disciplines, fostering communication and empathy skills while addressing ethical considerations in technology development. Ultimately, ASTRA 2025 aimed to inspire research engagement and innovation, equipping participants with the tools to contribute responsibly to the fields of neuroscience and AI.

<https://wpcms.computer.org/publications/tech-news/insider-membership-news/astra-neuroimaging-bci-ai/>

5. CFP: Feral Intelligence (FI): New Queer Approaches to Generative Artificial Intelligence (GAI)

Category: Policy & Ethics

Departments: Political Science & Public Administration, Sociology & Anthropology, Computer Science, Educational Leadership

Topics: Generative AI, AI Empire, Data Resurgence, Algorithmic Justice

Impact Score: 26.0

Rubric Scores

Research Relevance: 4

Educational Value: 4

Innovation / Novelty: 5

Practical Impact: 3

Interdisciplinary Potential: 5

Ethical / Policy Implications: 5

The article critiques Generative Artificial Intelligence (GAI) as a technology that perpetuates colonial and racial injustices, arguing that it is built on extracted data and exploitative labor practices. It highlights the concept of "AI Empire," which embodies the structural violence of colonialism and racial capitalism through algorithmic systems. The authors emphasize the importance of recognizing the harms caused by AI as inherent to these systems, while also acknowledging the resilience and innovative responses of marginalized communities. They advocate for a "data resurgence" that challenges the foundational assumptions of AI and promotes justice in the context of its impacts.

<https://digitalhumanitiesnow.org/2026/03/cfp-feral-intelligence-fi-new-queer-approaches-to-generative-artificial-intelligence-gai/>

6. Beyond the algorithm: rethinking the network account of trustworthy ai through lexical threshold-based multidimensional utility analysis

Category: Policy & Ethics

Departments: Computer Science, Political Science & Public Administration

Topics: AI Ethics & Safety, Trustworthiness Assessment, Culturally Distinct Trust, Network Account of Trust, Responsible AI Development

Impact Score: 25.0

Rubric Scores

Research Relevance: 4

Educational Value: 4

Innovation / Novelty: 5

Practical Impact: 3

Interdisciplinary Potential: 4

Ethical / Policy Implications: 5

Song's network account of trustworthy AI presents a framework that integrates various conceptions of trust, including both obligation-based and goodwill-based perspectives, while accommodating culturally distinct notions of trust. The framework posits that while AI technology lacks moral agency and can only be reliable, the broader AI network can be deemed trustworthy if AI companies act responsibly and ethical standards are maintained. The paper aims to enhance this network account by detailing the specific attributes required for four key nodes-technology, developers, professionals, and socio-legal systems-and proposing a novel method for assessing the overall trustworthiness of an AI system, particularly in cases where attributes may conflict. However, challenges remain in establishing a clear method for evaluating the trustworthiness of the entire AI network, which could hinder effective decision-making.

<https://link.springer.com/article/10.1007/s00146-026-02909-x>

7. A review of Robophilosophy, edited by Johanna Seibt, Raul Hakli, and Marco Nørskov

Category: Policy & Ethics

Departments: Philosophy, Political Science & Public Administration, Computer Science

Topics: AI Ethics & Safety, Social Robotics Governance, Moral Status of Robots, Philosophical Expertise in AI Policy

Impact Score: 25.0

Rubric Scores

Research Relevance: 4

Educational Value: 4

Innovation / Novelty: 3

Practical Impact: 4

Interdisciplinary Potential: 5

Ethical / Policy Implications: 5

In the edited volume "Robophilosophy," Seibt, Hakli, and Nørskov explore the implications of social robotics within the context of applied philosophy and technology governance. They identify a "triple gridlock" that obstructs research-based policymaking and argue for the necessity of philosophical expertise in addressing the challenges posed by the integration of social robots into everyday life. The volume discusses the moral status of social robots and the complexities of their deployment in various societal settings, emphasizing that current decisions regarding their development may have lasting consequences. While it does not provide a specific policy framework, the work is positioned as essential for understanding the conceptual issues surrounding the governance of social robots.

<https://link.springer.com/article/10.1007/s00146-026-02957-3>

8. Reasoning models struggle to control their chains of thought, and that's good

Category: Research

Departments: Computer Science, Political Science & Public Administration

Topics: AI Ethics & Safety, CoT-Control Framework, Reasoning Model Controllability, Safe AI Behavior

Impact Score: 25.0

Rubric Scores

Research Relevance: 5

Educational Value: 4

Innovation / Novelty: 5

Practical Impact: 4

Interdisciplinary Potential: 3

Ethical / Policy Implications: 4

OpenAI has introduced CoT-Control, a framework aimed at enhancing the monitorability of reasoning models. The research indicates that these models often struggle to effectively manage their chains of thought, which poses challenges for ensuring safe AI behavior. The findings underscore the importance of developing mechanisms to improve the controllability of AI reasoning processes as a safety measure.

<https://openai.com/index/reasoning-models-chain-of-thought-controllability>

9. Call For Papers: Special Issue on One Decade of GDPR - Contributions to Data Protection and Future Challenges

Category: Policy & Ethics

Departments: Political Science & Public Administration, Computer Science, Engineering Education & Leadership

Topics: AI Policy & Regulation, Privacy-Enhancing Technologies, Data Protection by Design, Impact of Geopolitical Changes, Usability of Privacy Notices

Impact Score: 25.0

Rubric Scores

Research Relevance: 4

Educational Value: 3

Innovation / Novelty: 4

Practical Impact: 4

Interdisciplinary Potential: 5

Ethical / Policy Implications: 5

The article announces a Special Issue dedicated to the 10th Anniversary of the EU General Data Protection Regulation (GDPR), inviting submissions that critically assess its implementation and the effectiveness of other global data protection frameworks, such as the California Consumer Privacy Act (CCPA). It highlights the GDPR's role in enhancing data subject rights and accountability, while also noting ongoing challenges in practical application, particularly regarding Data Protection by Design and the usability of privacy notices. The article emphasizes the need for further development of privacy-enhancing technologies in light of emerging threats from advanced tracking and AI systems, and calls for research addressing the impact of geopolitical and technological changes on privacy principles. Submissions are due by December 1, 2026, with publication scheduled for July/August 2027.

<https://wpcms.computer.org/digital-library/magazines/sp/cfp-one-decade-gdpr-challenges>

10. 5 Essential Security Patterns for Robust Agentic AI

Category: Policy & Ethics

Departments: Computer Science, Electrical & Computer Engineering, Political Science & Public Administration

Topics: AI Security, Just-in-Time Tool Privileges, Execution Sandboxing, Immutable Reasoning Traces

Impact Score: 25.0

Rubric Scores

Research Relevance: 4

Educational Value: 5

Innovation / Novelty: 4

Practical Impact: 5

Interdisciplinary Potential: 3

Ethical / Policy Implications: 4

The article discusses five key security patterns essential for safeguarding agentic AI systems, which are autonomous software entities. These patterns include Just-in-Time Tool Privileges, which provide temporary access rights to limit potential damage if compromised; Bounded Autonomy, which allows agents to operate independently within defined parameters while requiring human oversight for sensitive actions; and the AI Firewall, which filters user inputs to prevent threats like prompt injection. Additionally, Execution Sandboxing isolates agent-generated code to mitigate risks, and Immutable Reasoning Traces create tamper-evident logs for auditing decisions, enhancing transparency and accountability in high-stakes applications. The article emphasizes that these security measures are most effective when implemented as a layered system.

<https://machinelearningmastery.com/5-essential-security-patterns-for-robust-agentic-ai/>