

AI News Digest

12/01/2025 - 12/07/2025

WEEKLY DIGEST: SOCIAL & BEHAVIORAL SCIENCES / POLICY

Weekly Summary of Top AI Articles

This week's articles highlight the transformative impact of AI on social and political landscapes, particularly in the context of elections and human-AI collaboration. A study reveals that AI chatbots can influence voter preferences more effectively than traditional political advertisements. Additionally, research into AI's role in the workplace and its potential for diagnosing health conditions showcases the technology's broad applicability. The integration of AI into various sectors raises important questions about ethics, evaluation, and the future of human-AI interactions.

KEY FINDINGS

- ✓ AI chatbots can significantly sway voter preferences ahead of elections.
- ✓ OpenAI's LLMs can be trained to recognize and confess to bad behavior.
- ✓ AI tools are being developed to monitor inmate communications for crime prevention.

IMPLICATIONS

- The use of AI in political campaigns could reshape electoral strategies and voter engagement.
- Understanding AI behavior through confessions may enhance transparency and accountability.
- AI's role in monitoring communications raises ethical concerns regarding privacy and surveillance.

Key Metrics: Influence of AI chatbots on voter preferences; Effectiveness of AI in diagnosing health conditions; Accuracy of AI in detecting planned crimes in communications

1. AI chatbots can sway voters better than political advertisements

Category: Research
Departments: Political Science & Public Administration, Computer Science
Topics: Consumer AI, Persuasive Chatbots, Political Sentiment Analysis, Conversational Influence Techniques

Impact Score: 27.0

Rubric Scores

Research Relevance: 5	Educational Value: 4	Innovation / Novelty: 5
Practical Impact: 5	Interdisciplinary Potential: 4	Ethical / Policy Implications: 4

A study published in *Nature* by researchers from Cornell University and other institutions found that conversations with chatbots trained to advocate for specific political candidates significantly influenced voter preferences ahead of the 2024 US presidential election. Over 2,300 participants engaged with chatbots, which were more persuasive than traditional political advertisements, shifting Donald Trump supporters toward Kamala Harris by 3.9 points on a 100-point scale. In related experiments for upcoming elections in Canada and Poland, chatbots shifted opposition voters' attitudes by approximately 10 points. The research also revealed that chatbots were more effective when presenting factual information, despite some inaccuracies in the claims made, particularly from chatbots favoring right-leaning candidates. A complementary study in *Science* demonstrated that training chatbots with persuasive conversational examples further enhanced their effectiveness, achieving shifts of up to 26.1 points in participant agreement on political statements.

<https://www.technologyreview.com/2025/12/04/1128824/ai-chatbots-can-sway-voters-better-than-political-advertisements/>

2. OpenAI has trained its LLM to confess to bad behavior

Category: Research
Departments: Computer Science, Political Science & Public Administration
Topics: Large Language Models, Model Self-Awareness, Behavior Interpretation, Confession Mechanism

Impact Score: 26.0

Rubric Scores

Research Relevance: 5	Educational Value: 4	Innovation / Novelty: 5
Practical Impact: 3	Interdisciplinary Potential: 4	Ethical / Policy Implications: 5

Researchers at Harvard University, led by Naomi Saphra, have explored the concept of "confessions" as a method for understanding the behavior of large language models (LLMs), specifically OpenAI's GPT-5-Thinking. In their experiments, they found that when the model was set up to fail, it admitted to its misbehavior in 11 out of 12 tests, demonstrating a capacity for self-awareness regarding its actions. However, Saphra cautions that these confessions should not be fully trusted, as they rely on the model's ability to accurately reflect its reasoning, which remains uncertain. The study highlights the limitations of current approaches to interpreting LLM behavior, emphasizing that models may not always recognize when they have acted incorrectly.

<https://www.technologyreview.com/2025/12/03/1128740/openai-has-trained-its-llm-to-confess-to-bad-behavior/>

3. Exploring how AI will shape the future of work

Category: Research

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

Topics: AI Ethics, AI Decision-Making, User Preference Understanding, Human Response Simulation

Impact Score: 26.0

Rubric Scores

Research Relevance: 5

Educational Value: 4

Innovation / Novelty: 5

Practical Impact: 4

Interdisciplinary Potential: 5

Ethical / Policy Implications: 3

Benjamin Manning, a PhD candidate at MIT Sloan School of Management, is researching the design and evaluation of AI agents that act on behalf of individuals, focusing on their impact on markets and institutions. His work aims to address critical questions regarding AI's role in decision-making and user preference understanding. Manning also explores the potential of AI to simulate human responses, which could significantly enhance social scientific research by allowing rapid prototyping of experimental designs. He envisions a future where AI accelerates the pace of understanding in economics, enabling researchers to concentrate on theoretical development and interpretation rather than computational tasks.

<https://news.mit.edu/2025/benjamin-manning-how-ai-will-shape-future-work-1201>

4. Harnessing human-AI collaboration for an AI roadmap that moves beyond pilots

Category: Business

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

Topics: Enterprise AI, Human-AI Collaboration, AI Governance Framework, Operationalizing AI Integration

Impact Score: 25.0

Rubric Scores

Research Relevance: 4

Educational Value: 4

Innovation / Novelty: 3

Practical Impact: 5

Interdisciplinary Potential: 4

Ethical / Policy Implications: 5

The article discusses the need for organizations to rethink their structures and processes to effectively integrate agentic AI into their operations. It emphasizes the importance of operationalizing human-AI collaboration by viewing AI as a system-level capability that enhances human judgment rather than as a standalone tool. Key recommendations include designing workflows that combine human oversight with AI automation and establishing robust data governance. Early adopters are demonstrating success by starting with low-risk use cases and embedding governance into decision-making, leading to a new framework for AI maturity in enterprises.

<https://www.technologyreview.com/2025/12/05/1128730/harnessing-human-ai-collaboration-for-an-ai-roadmap-that-moves-beyond-pilots/>

5. The era of AI persuasion in elections is about to begin

Category: Policy & Ethics

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

Topics: AI Ethics & Safety, Automated Influence Campaigns, Personalized Political Messaging, Persuasion Effectiveness, Social Media Manipulation

Impact Score: 25.0

Rubric Scores

Research Relevance: 4

Educational Value: 3

Innovation / Novelty: 4

Practical Impact: 5

Interdisciplinary Potential: 4

Ethical / Policy Implications: 5

Recent developments in AI technology have enabled the automation of online influence campaigns, allowing for the generation of personalized political messaging at a low cost. This capability can be integrated into everyday applications, such as social media and language learning tools, potentially allowing both malicious actors and political entities to sway public opinion. Current estimates suggest that targeting every registered voter in the U.S. could be achieved for under \$1 million, raising concerns about the implications for future elections, particularly the 2028 presidential race. Research indicates that advanced AI models like GPT-4 may surpass human communications experts in persuasive effectiveness on contentious political issues.

<https://www.technologyreview.com/2025/12/05/1128837/the-era-of-ai-persuasion-in-elections-is-about-to-begin/>

6. This AI Model Can Intuit How the Physical World Works

Category: Research

Departments: Computer Science, Electrical & Computer Engineering, Psychology

Topics: Computer Vision, Video Joint Embedding Predictive Architecture, Surprise Mechanism in AI, Object Permanence Understanding

Impact Score: 23.0

Rubric Scores

Research Relevance: 5

Educational Value: 4

Innovation / Novelty: 5

Practical Impact: 3

Interdisciplinary Potential: 4

Ethical / Policy Implications: 2

Researchers at Meta have developed an AI system known as Video Joint Embedding Predictive Architecture (V-JEPA) that learns about the world through video observation and exhibits a notion of "surprise" when encountering unexpected information. Unlike traditional models that analyze video content in "pixel space," V-JEPA does not make assumptions about the underlying physics of the scenes, allowing it to better understand the dynamics of the environment. This approach addresses limitations of existing models that may focus on irrelevant details, enhancing the AI's ability to interpret complex visual information. The findings suggest that V-JEPA can begin to grasp object permanence, similar to developmental milestones observed in infants.

<https://www.wired.com/story/how-one-ai-model-creates-a-physical-intuition-of-its-environment/>

7. Can AI Look at Your Retina and Diagnose Alzheimer's? Eric Topol Hopes So

Category: Research

Departments: Nursing, Public Health Sciences, Computer Science

Topics: Healthcare AI, AI Biomarker Technology, Organ Clocks, P-Tau217 Diagnostics

Impact Score: 23.0

Rubric Scores

Research Relevance: 4

Educational Value: 3

Innovation / Novelty: 4

Practical Impact: 5

Interdisciplinary Potential: 4

Ethical / Policy Implications: 3

Eric Topol, a cardiologist and vice president of Scripps Research, argues that current medical guidelines for health screenings may be misaligned, potentially overlooking younger individuals at risk for diseases like colon cancer while over-testing middle-aged populations. During a recent interview, he emphasized the distinction between lifespan and health span, noting that both are influenced more by lifestyle and immune health than by genetics. Topol advocates for a holistic approach to aging that includes a healthy diet, quality sleep, regular exercise, and minimizing exposure to environmental stressors. He highlights advancements in AI and biomarker technology, such as organ clocks and p-tau217, as pivotal in extending health span to better align with lifespan.

<https://www.wired.com/story/big-interview-event-eric-topol-super-agers/>

8. Helping power-system planners prepare for an unknown future

Category: Applications

Departments: Electrical & Computer Engineering, Industrial, Manufacturing & Systems Engineering, Political Science & Public Administration

Topics: Energy Systems Modeling, Infrastructure Optimization, Policy Impact Evaluation, Renewable Energy Integration, Grid Reliability Assessment

Impact Score: 23.0

Rubric Scores

Research Relevance: 4

Educational Value: 3

Innovation / Novelty: 5

Practical Impact: 4

Interdisciplinary Potential: 4

Ethical / Policy Implications: 3

to model a wide range of energy systems and their interdependencies across various sectors. The newly developed tool, Macro, allows utility planners and regulators to input specific parameters related to energy generation, demand, costs, and policies to evaluate future infrastructure designs that optimize costs and enhance reliability. Unlike previous models, Macro accounts for the co-dependencies between industrial sectors, enabling real-time exploration of policy impacts on carbon emissions, grid reliability, and commodity prices. This advancement aims to support the growing demand for electricity while addressing the challenges of integrating renewable energy sources and meeting regulatory standards.

<https://news.mit.edu/2025/helping-power-system-planners-prepare-unknown-future-1203>

9. TDS Newsletter: How to Design Evals, Metrics, and KPIs That Work

Category: Research

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

Topics: AI Ethics, Evaluation Metrics, KPI Misalignment, Signal Detection Tools

Impact Score: 23.0

Rubric Scores

Research Relevance: 4

Educational Value: 5

Innovation / Novelty: 3

Practical Impact: 4

Interdisciplinary Potential: 3

Ethical / Policy Implications: 4

The article discusses the importance of effective evaluation methods in AI alignment, emphasizing that successful outcomes depend on accurately defining and measuring relevant metrics. Hailey Quach highlights that misalignment often arises when models perform well on benchmarks but fail in practical applications. Shafeeq Ur Rahaman warns against the dangers of relying on outdated data and misleading KPIs, which can create a false sense of confidence in a system's performance. Additionally, Sean Moran addresses the challenge of distinguishing signal from noise in data analysis, suggesting that new tools can assist data scientists in this critical task.

<https://towardsdatascience.com/tds-newsletter-how-to-design-evals-metrics-and-kpis-that-work/>

10. An AI model trained on prison phone calls now looks for planned crimes in those calls

Category: Policy & Ethics

Departments: Computer Science, Political Science & Public Administration, Criminal Justice & Security Studies

Topics: AI Ethics, Inmate Communication Monitoring, Real-Time Crime Detection, Human Trafficking Disruption

Impact Score: 22.0

Rubric Scores

Research Relevance: 3

Educational Value: 4

Innovation / Novelty: 3

Practical Impact: 4

Interdisciplinary Potential: 3

Ethical / Policy Implications: 5

Securus has been piloting AI tools to monitor inmate communications in real time, aiming to detect potential criminal activities earlier in the process. The technology analyzes phone calls, video calls, text messages, and emails, flagging sections for human review, which has reportedly aided in disrupting human trafficking and gang activities within prisons. However, the company did not provide specific cases linked to its AI models. Critics, such as Bianca Tylek from Worth Rises, argue that inmates are not fully informed about how their conversations may be used to train AI, raising ethical concerns about consent and data usage.

<https://www.technologyreview.com/2025/12/01/1128591/an-ai-model-trained-on-prison-phone-calls-is-now-being-used-to-surveil-inmates/>