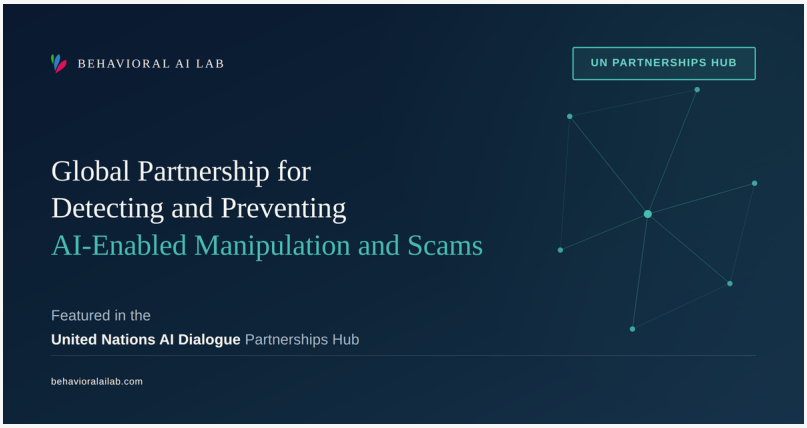
Behavioral AI Lab Launches UN-Recognized International AI Safety Initiative to Combat AI-Enabled Manipulation and Scams

New international initiative advances behavioral AI evaluation to detect AI-enabled manipulation, online scams, and emerging AI safety risks.



Behavioral AI Lab launches the Global Partnership for Detecting and Preventing AI-Enabled Manipulation and Scams, featured in the United Nations Global Dialogue on AI Governance Partnerships Hub.

SAN FRANCISCO, CA, UNITED STATES, August 5, 2026 /EINPresswire.com/ -- Behavioral AI Lab today announced the launch of its international AI safety initiative, the Global Partnership for Detecting and Preventing AI-Enabled Manipulation and Scams. The initiative is listed in the AI Dialogue Partnerships Hub, part of the United Nations Global Dialogue on AI Governance, with Behavioral AI Lab serving as the Implementing Entity.

Artificial intelligence is rapidly transforming how people communicate, make decisions, and navigate digital environments. While these advances create enormous opportunities, they also enable increasingly sophisticated forms of psychological manipulation, fraud, and deception that traditional safety systems struggle to detect.

> Behavioral science provides the missing human layer for building safer, more trustworthy AI systems."
>
> *Tomomi Tanaka, Founder, Behavioral AI Lab*

Rather than relying solely on keyword filtering or reactive moderation, the Lab develops behavioral evaluation methods that identify the psychological and behavioral patterns underlying manipulation. Its goal is to help AI developers, Trust & Safety professionals, governments, and technology companies build systems that prevent harm before it occurs.

The partnership's first major research focus is AI-enabled manipulation and scams—including

cryptocurrency investment fraud, romance scams, and impersonation attacks—but its broader mission is to advance AI safety through human-centered evaluation and Safety by Design.

According to the FBI's Internet Crime Complaint Center (IC3), losses related to cryptocurrency fraud exceeded USD 11 billion in 2025, including USD 7.2 billion from cryptocurrency investment fraud alone.

The initiative builds on the Lab's research analyzing more than 15,000 scammer messages collected from more than 150 documented cryptocurrency romance scam cases across multiple languages and cultural contexts. The research demonstrates that focusing on the psychology of persuasion and manipulation—rather than surface-level keywords—provides a more robust and generalizable approach to detecting increasingly adaptive forms of AI-enabled deception.

A Global Partnership for Safer AI

The partnership brings together experts in behavioral science, AI governance, Trust & Safety, and multilingual AI development. The Lab leads the partnership as the Implementing Entity.

Participating organizations include Shisa.AI, a Tokyo-based AI company specializing in multilingual Small Language Models (SLMs). Shisa.AI contributes multilingual SLM development, safety-focused post-training methods, dataset development, and automated model evaluation. Additional research and implementation partners are expected to join as the initiative expands.

The partnership will develop practical open resources for researchers, AI developers, Trust & Safety professionals, financial institutions, consumer protection agencies, and policymakers, including:

- Behavioral Manipulation Taxonomy v1 and implementation guidance
- Multilingual behavioral-risk datasets
- Behavioral AI evaluation benchmarks
- Safety by Design framework for AI-enabled manipulation prevention
- Capacity-building workshops across at least three countries
- Pilot implementations with global partner organizations
- A multistakeholder knowledge-sharing network involving at least ten organizations

Bringing Behavioral Science into AI Safety

"AI safety is no longer just about preventing harmful outputs. As AI becomes increasingly capable of influencing human decisions, we must understand how people decide whom to trust, how cognitive biases are exploited, and how manipulation develops over time.

Behavioral science provides the missing human layer in today's AI safety ecosystem. By integrating behavioral science into AI evaluation, we aim to help build AI systems that are not

only more capable, but also more trustworthy and better aligned with human well-being."

— Tomomi Tanaka, Founder of Behavioral AI Lab

Tomomi Tanaka is the founder of Behavioral AI Lab and a behavioral scientist and AI strategist specializing in AI safety, behavioral science, and Trust & Safety. Before founding Behavioral AI Lab, she held leadership positions at the World Bank, Uber, Match Group, Amazon, and Oracle. Her research has appeared in the American Economic Review, and her writing has been published in the Harvard Business Review. She has spoken internationally on AI governance, AI evaluation, and Safety by Design, including as an invited speaker at United Nations Headquarters in New York.

About Behavioral AI Lab

Behavioral AI Lab is an independent research and advisory organization based in San Francisco. The organization combines behavioral science and artificial intelligence to improve the safety, trustworthiness, and human-centered design of AI systems. Its work spans AI evaluation, Safety by Design, Trust & Safety, human-AI interaction, behavioral manipulation, scam prevention, and responsible AI governance. For more information, visit https://behavioralailab.com

Tomomi Tanaka
Behavioral AI Lab
+1 424-438-6173
tomomit@behavioralailab.com
Visit us on social media:
LinkedIn
Bluesky
YouTube
X
Other
