AI SAFETY

EPISTEMIC SECURITY

CRITICAL COGNITIVE INFRASTRUCTURE

A FRAMEWORK WITHIN THE EPISTEMIC SECURITY FIELD

BY KAVYA PEARLMAN

XRSI™
Human Intelligence In The Loop

15 SEP 2026

CRITICAL COGNITIVE INFRASTRUCTURE

A FRAMEWORK WITHIN THE EPISTEMIC SECURITY FIELD

SUMMARY

Epistemic security is an established field, researched within frontier AI governance as well as UK democracy and media policy. Critical Cognitive Infrastructure (CCI) is a specific, operational framework within it, defining the verified information channels that democratic institutions depend on as a distinct, regulable object. As transformative AI capability accelerates, protecting this infrastructure is the precondition for AI safety, not an adjacent concern, since oversight and course-correction depend on institutions retaining the capacity to reason clearly. Closing this gap requires a comprehensive effort, not whack-a-mole responses to individual incidents: coordinated research, a practical toolkit, certifiable standards, and a federal governance plan.

1. Epistemic security is an established field, including inside frontier AI governance

Epistemic security describes the effort to secure the information supply chains that democratic decision-making depends on. The term originates in Seger et al. (2020), published by the Alan Turing Institute, which framed threats to informed decision-making in democratic societies as a security problem, not only a media or content problem.[1]

Demos, a UK cross-party think tank, has built a dedicated Epistemic Security Network with Seger as a co-author of its reports. The first, Epistemic Security 2029, was published in February 2025.[2] The second, Epistemic Security: Information Resilience Priorities for 2026-2029, was published in August 2026.[3] This body of work centers on the UK's information supply chain: news ecosystems, public service broadcasting, platform regulation, and electoral law, and treats large language models and AI chatbots as one threat vector among several.

2. Where the field's current work sits, and where CCI differs

The existing work, both Seger's research at GovAI and the Demos reports, is academic and policy research: analysis, cause-area assessment, and legislative recommendation. It has not, as far as this review has found, produced certifiable standards, an operational federal governance proposal, or field-building infrastructure connecting AI safety practitioners to the governance and standards world. CCI's contribution is that operational layer: a specific framework built for certification and evaluation, a proposal addressed to US federal agencies rather than UK Parliament, and a convening effort connecting practitioners across these communities. This is a narrower and more precise claim than asserting the field itself is disconnected from AI safety, and it should be verified against the fuller body of Seger's and GovAI's published work before being relied upon.

CISA itself has already engaged this framing directly. Then-Director Jen Easterly stated publicly, at WIRED's RE:WIRED conference on November 10, 2021, that cognitive infrastructure is arguably the nation's most critical infrastructure, and that building resilience to misinformation and disinformation is essential to CISA's mission.[4] The following month, CISA's own Cybersecurity Advisory Committee established a subcommittee, "Protecting Critical Infrastructure from Mis-, Dis-, and Mal-information," with a mandate to evaluate the agency's role in addressing harms to election systems, telecommunications, and public health infrastructure.[5] CCI's proposal to CISA builds on and formalizes work the agency's own advisory structure already began exploring, rather than proposing an entirely new mandate.

XRSI's own engagement with this framing predates the current memo. A roundtable report from Metaverse Safety Week 2023, published by XRSI in January 2024, explicitly called for "Strengthening Cognitive Critical Infrastructure" in the context of AI and immersive technology convergence, citing earlier work on national cognitive infrastructure protection.[6][7]

3. The threat is active now, not projected

Three independent sources confirm this is a current, scaling risk rather than a future one.

Anthropic's September 2026 threat intelligence report documents AI-enabled election-targeting operations across multiple countries (Moldova, Malaysia, Kenya, the Central African Republic), attribution-laundering techniques designed to manufacture false certainty, and impersonation of real institutions and officials, including forged government documents and fabricated UN testimony. Most cases were caught through voluntary company detection or civil society tips, not government mechanisms.[8]

OpenAI's threat reports from February and June 2026 independently confirm the same pattern, disrupting China- and Russia-linked influence operations and PRC-linked clusters using its models to shape US AI policy debates.[9] Two competing AI companies observing the same behavior independently is stronger evidence than either report alone.

The World Economic Forum's Global Risks Report 2026, surveying more than 1,300 global leaders and experts, ranks misinformation and disinformation the second most severe global risk over the next two years. Adverse outcomes of AI is the fastest-rising risk the report has tracked, moving from thirtieth to fifth place between the two-year and ten-year outlooks.[10]

"The most dangerous manipulation is the kind you never feel happening, a quiet erosion of your own judgment. Epistemic security is the fight against that erosion. Critical Cognitive Infrastructure names what we are protecting: the channels of truth a society needs to think clearly, before it forgets how."

– Kavya Pearlman, Founder & Principal Researher, XRSI

4. Critical Cognitive Infrastructure: the framework

CCI narrows epistemic security to a specific, regulable object: the verified channels through which government, electoral, public health, and judicial bodies transmit information to the public, where compromise of the channel produces judgment failure at population scale.

Working definition: the framework designates and regulates these channels. It does not designate or regulate the cognitive capacity of the individuals who rely on them. Protecting that capacity is the outcome. The channel is the regulated object.

This differs from the existing response landscape in scope and mechanism. C2PA and similar content-provenance standards verify whether a piece of content is authentic. They do not govern institutional resilience or graduated response. NIST's guidance on synthetic content provenance is voluntary. State laws, including California's AI Transparency Act and New York's synthetic-performer disclosure law, are fragmented and inconsistent across jurisdictions. Lawmakers pressing AI companies directly for election-integrity commitments have received no binding response, and have stated publicly that the scale of the problem demands more coordinated action than currently exists.

CCI proposes three governance actions, detailed in a companion memorandum: designating Critical Cognitive Infrastructure as a protected category with a defined threat assessment process; developing graduated risk thresholds and resilience standards, building on the NIST AI Risk Management Framework; and establishing a coordinating body across the agencies currently addressing this problem in isolation.

XRSI™
Human Intelligence In The Loop

5. Positioning

XRSI has built safety and governance standards for AI and immersive technology since 2020, including the Responsible Data Governance standard, and advises more than 60 governments and international bodies including the OECD and the UN International Telecommunication Union. Prior to founding XRSI, its founder advised Facebook on third-party security risks during the 2016 US presidential election.

This work matters now because of what transformative AI does to the precondition for every other safeguard. Oversight, democratic correction, and course-correction all depend on one thing: people and institutions retaining the capacity to reason clearly and detect when they are being misled. The evidence in this paper shows that capacity is already being eroded at scale, and the AI Futures Project's own scenario for the years ahead depicts propaganda operations and public confusion compounding as capability accelerates, without treating this as a governance target in its own right.[11] If human judgment degrades faster than institutions can adapt, technical alignment work and model evaluation will not be enough, because the people meant to catch a failure will not be able to see it clearly. Protecting cognitive infrastructure is not adjacent to AI safety. It is the foundation the rest of AI safety depends on.

CCI extends XRSI's track record into this specific problem. Epistemic security already connects to frontier AI governance research, as this paper's citations show. What is missing is the operational layer: certifiable standards, a federal governance proposal, and practitioner infrastructure that turns this research into something institutions can implement before the transition accelerates further. Protecting Critical Cognitive Infrastructure requires a comprehensive effort, not whack-a-mole responses to individual incidents: coordinated research, a practical toolkit, policy, and a plan that ties them together.

References

1. Seger, E., Avin, S., Pearson, G., Briers, M., Ó hÉigeartaigh, S., Bacon, H. (2020) Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically advanced world. The Alan Turing Institute.
2. Seger, E., Hancock, J., Perry, H. (2025) Epistemic Security 2029: Fortifying the UK's information supply chain to tackle the democratic emergency. Demos, February 2025.
3. Perry, H., Moores, A., Devine, F. (2026) Epistemic Security: Information Resilience Priorities for 2026-2029. Demos, August 2026.
4. Ferris, S. (2021, November 11) Cyber agency beefing up disinformation, misinformation team. The Hill, reporting remarks by CISA Director Jen Easterly at WIRED's RE:WIRED conference, November 10, 2021.
5. CISA Cybersecurity Advisory Committee (2021, December) Protecting Critical Infrastructure from Mis-, Dis-, and Mal-information, subcommittee mission statement, as reported in The CISA CSAC: Cognitive Infrastructure Research and Election Public Messaging, OODA Loop, October 2022.
6. XRSI (2024, January) Navigating the Cyber Frontier in the AI-Powered Metaverse: A Roundtable Report. Metaverse Safety Week 2023.
7. Pereira, D. (2022, January 4) National Cognitive Infrastructure Protection: What Can We Learn from the Swedish Psychological Defence Authority? OODA Loop.
8. Anthropic (2026) Threat Intelligence Report, September 2026.
9. OpenAI (2026) Influence and cyber operations threat reports, Disrupting malicious uses of AI, February 2026 and PRC-linked influence operations are targeting AI debates in the US, June 2026.
10. World Economic Forum (2026) Global Risks Report 2026.
11. Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., Dean, R. (2025) AI 2027. AI Futures Project. Published April 3, 2025.

XRSI™
Human Intelligence In The Loop